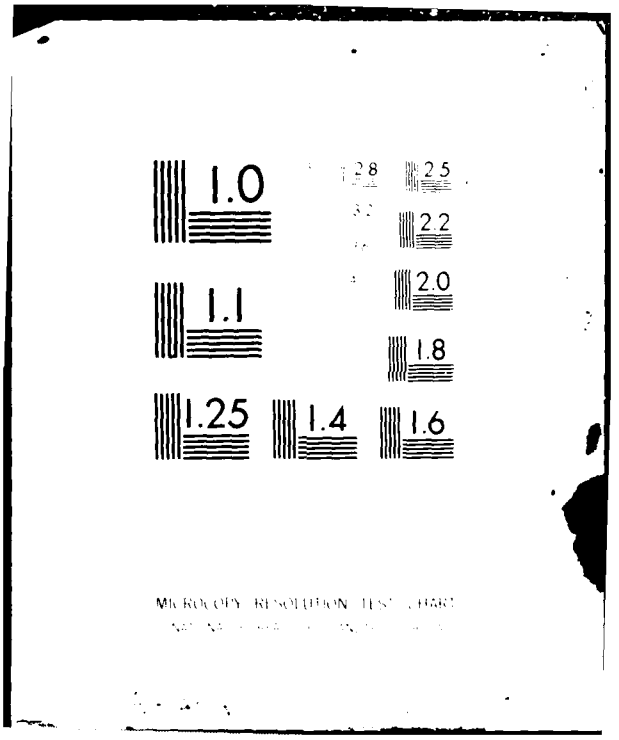
AD-A116 344 NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 5/9
SYMPOSIUM PROCEEDINGS: PRODUCTIVITY ENHANCEMENT: PERSONNEL PERF--ETC(U)
1977 L T POPE, D MEISTER

UNCLASSIFIED NL

1 OF 5
AD A
116 344


1.0
1.1
1.25
1.4
1.6
2.0
1.8
2.2
2.5
MICROCOPY RESOLUTION TEST CHART

AD A116344

DTIC FILE COPY

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

SYMPOSIUM PROCEEDINGS

PRODUCTIVITY ENHANCEMENT: PERSONNEL PERFORMANCE ASSESSMENT IN NAVY SYSTEMS

OCTOBER 12 - 14, 1977

DTIC ELECTE JUN 29 1982 S D A

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

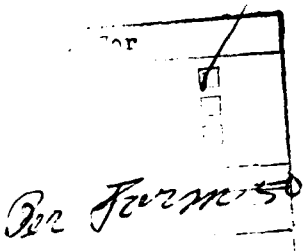
82 06 28 118

SYMPOSIUM PROCEEDINGS:

PRODUCTIVITY ENHANCEMENT: PERSONNEL PERFORMANCE ASSESSMENT IN NAVY SYSTEMS

October 12-14, 1977

Edited by
Louis T. Pope
and
David Meister



Reviewed by
Frederick A. Muckler

Navy Personnel Research and Development Center
San Diego, California 92152

FOREWARD

The proceedings and papers that comprise this report have been prepared as part of Project Z0107-PN.05, Personnel Performance Capabilities, the purpose of which is to develop more effective methods of measuring personnel performance in the Navy's operational environment. It is felt that the symposium--entitled, "Productivity Enhancement: Personnel Performance Assessment in Navy Systems," to be held 12-14 October 1977--will contribute materially to the goal of this project by bringing together specialists from the three military services and from the civilian sector to review the status of performance measurement methodology and to recommend further research in this area. The opinions expressed in these papers are those of the authors and do not necessarily represent the views of the Navy Personnel Research and Development Center or the U.S. Navy.

Many individuals have contributed to the preparation of this Proceedings and to completing the countless details associated with this technical Symposium. We wish in particular to thank the clerical staff of the NPRDC Design of Manned Systems Program: Ms. Evelyn Wilson, Ms. Carolyn Shaw, and Mr. Stephen Eastburn.

In the military,
" things are good, ideas are bad"
Clarkin

PRECEDING PAGE BLANK-NOT FILMED

CONTENTS

	Page
INTRODUCTION--Louis T. Pope	1
WHAT DOES PERFORMANCE MEASUREMENT MEAN? (Abstract) Dr. E. A. Fleishman and Dr. J. M. Levine	3
AIRCREW PERFORMANCE MEASUREMENT--D. Vreuls and L. Wooldridge	5
PERFORMANCE MEASUREMENT OF SHIPBOARD OPERATIONAL SKILLS A. Anderson and E. Pickering	33
PERFORMANCE MEASUREMENT OF MAINTENANCE--Dr. J. P. Foley	55
SOME PROBLEMS IN TEAM PERFORMANCE--Dr. J. J. Collins	83
PERFORMANCE MEASUREMENT IN SYSTEM TEST AND EVALUATION LCDR W. Moroney, USN and LT W. Helm, USN	91
THE CHARACTERISTICS OF NAVAL PERSONNEL AND PERSONNEL PERFORMANCE S. A. Horowitz and LCDR A. Sherman, USN	109
PERFORMANCE MEASUREMENT IN CIVILIAN ORGANIZATIONS: APPLICATIONS TO THE MILITARY SETTING--Dr. M. Sanders	123
PERFORMANCE MEASUREMENT SYSTEM ARCHITECTURE AND DATA PROCESSING LOADS--R. W. Obermayer	135
AUTOMATION OF PERFORMANCE MEASUREMENT--Dr. R. C. Williges	153
SELECTING MEASURES: "OBJECTIVE" VS "SUBJECTIVE" MEASUREMENT Dr. F. A. Muckler	169
USING SIMULATORS FOR PERFORMANCE MEASUREMENT A. Crawford and J. F. Brock	179
MEASUREMENT OF PRODUCTIVITY ENHANCEMENT: EVALUATING A PERFORMANCE-CONTINGENT REWARD SYSTEM THAT USES ECONOMIC INCENTIVES Dr. G. E. Bretton, Dr. S. L. Dockstader, Dr. D. M. Nebeker, and Dr. E. C. Shumate	193
EFFECTS OF THE OPERATIONAL ENVIRONMENT ON PERFORMANCE CAPT J. J. Clarkin, USN	249
THE HUMAN SIDE OF PERFORMANCE MEASUREMENT--Dr. L. Broedling	261
THE STRATEGY OF PERFORMANCE MEASUREMENT--Dr. D. Meister	277
PERFORMANCE TESTING IN INSTRUCTIONAL SYSTEMS--J. F. Brock	303
ON THE MEASUREMENT OF MAN-MACHINE PERFORMANCE--Dr. R. Mackie	323

PRECEDING PAGE BLANK-NOT FILMED

Page

PERFORMANCE MEASUREMENT TECHNOLOGY: ISSUES AND ANSWERS
Dr. E. A. Alluisi 343

PLANNING FOR PERFORMANCE MEASUREMENT R&D: U. S. ARMY
Dr. M. Katz 361

PLANNING FOR AIRCREW PERFORMANCE MEASUREMENT R&D: U. S. AIR FORCE
Dr. W. Waag and Ms. P. Knoop 381

THE DEVELOPMENT OF A NAVY PERFORMANCE EVALUATION TEST FOR ENVIRONMENTAL RESEARCH (PETER)--CDR R. S. Kennedy, USN and A. C. Bittner, Jr. 393

INTRODUCTION

Louis T. Pope
Navy Personnel Research and Development Center
San Diego, California

The ability to measure personnel performance validly and reliably is central to the program goals of the Navy Personnel Research and Development Center and, indeed, to all research organizations dedicated to improving man-machine relationships in military systems. It is only a slight exaggeration to say that a phenomenon that cannot be measured adequately cannot be improved (or at least deliberately improved). Hence the need to develop more effective means of assessing personnel performance, particularly in the context of the operational environment (i.e., at sea and on shore in the performance of mission-related tasks). As the Navy's technological sophistication increases, the complexity of these tasks also increases and makes performance measurement more difficult.

Personnel performance assessment is involved in many aspects of Navy missions: (1) in the selection of appropriate personnel for demanding jobs, (2) in the test and evaluation of new weapons and their support systems, (3) in the determination of ship operational readiness for combat, and (4) in the assessment of individual capability for assignment and promotion. The methodology presently employed in these assessments demands considerable improvement because it rests too heavily on a subjective "expertise" that is all too often neither expert nor methodologically sound.

The emphasis in this symposium has been on the context that presents the greatest difficulty for measurement, that of individuals and crews performing complex tasks in operational missions. Papers and workshops illustrating the need for and value of performance measurement in the enhancement of productivity have also been included.

The purpose of the presentations is to help:

1. To determine the status of existing techniques and present and propose research on personnel performance measurement.

2. To facilitate the exchange of information on these topics between researchers and operational personnel.

3. To stimulate the generation of new approaches in setting goals and in making plans for personnel performance research.

It is entirelv appropriate that the three services are represented in this symposium because all of them face similar measurement problems. It is also appropriate that the civilian community is represented because it is deeply involved in research to help solve these problems.

The Navy Personnel Research and Development Center is indebted to the many researchers who agreed to present papers, and to RADM William R. Smedberg and Ms. Mitzi Wertheim for their continuing support. We must also extend our appreciation to the attendees who, by their active participation in the symposium's discussions, are ensuring continued progress in improving personnel performance assessment methodology.

WHAT DOES PERFORMANCE MEASUREMENT MEAN?

Edwin A. Fleishmen and Jerrold M. Levine
Advanced Research Resources Organization
Silver Spring, Maryland

ABSTRACT

A general introduction to performance measurement is presented which centers around the answer to two key questions: How do we know what to measure? How do we generalize what we already know?

With regard to the first question, issues dealing with taxonomies or classification systems are discussed, types of categorization systems previously examined are reviewed, and criteria for evaluating these systems are suggested. Other complexities and factors to be considered at the beginning of the process of measuring performance in systems are also pointed out.

With regard to the second question, the need for, uses of, and techniques for development of a human performance data base are discussed. This is viewed as an end point in the measurement process which allows for the integration of empirical data across studies in order to predict performance on new systems as they are developed. The use of computerized information retrieval systems for storing and accessing this data base is also discussed.

ABOUT THE AUTHORS

Dr. Edwin A. Fleishman is President of the Advanced Research Resources Organization. He has been a professor at Yale University and the University of California, and was formerly Director of the American Institutes for Research in Washington. He is the author of books and articles on human abilities and performance measurement. From 1971 through 1976 he was editor of the Journal of Applied Psychology. He is Past President of the American Psychological Association's Division of Industrial and Organizational Psychology, is current President of the Division of Engineering Psychology and is President Elect of the Division of Evaluation and Measurement. He is also currently serving an 8 year term as President of the International Association of Applied Psychology. In 1974, he received the Franklin Taylor award for contributions to engineering psychology.

Dr. Jerrold M. Levine is a member of the scientific staff of the Advanced Research Resources Organization where he is engaged in and directs research on the assessment of human performance, on the effects of alcohol and drugs on behavior, on human factors in command and control systems, and on highway safety problems. Formerly he was Director of Behavioral Science at Science Applications, Inc. and Director of Human Performance Research at the American Institutes for Research. Previous to that he was affiliated with Rockwell Industries. He received his Ph.D. from the University of Massachusetts in 1966.

AIRCREW PERFORMANCE MEASUREMENT

Donald Vreuls and Lee Wooldridge
Canyon Research Group, Inc.
Westlake Village, California

ABSTRACT

Air crew performance measurement is described in terms of the aircrew environment, an approach to measurement development, and future research needs. A global view of the aircrew environment barely touches a few of the variables and considerations involved in the training process, the aircraft and weapon system environment, and the operational environment. One approach to measurement development is used to examine some of the considerations, progress, and methodological issues in selected areas of analysis for measurement, measurement system design, data collection, measure selection techniques, and product and system effectiveness testing. Future research needs for more empirical data, better analytic methods, measurement standardization, and personnel are highlighted. Aircrew performance measurement has come a long way in the past several years, but there is much more to do if we are going to be fully responsive to growing needs.

INTRODUCTION

The role of measurement is to provide information needed to guide decisions by policy makers, technology managers, engineers, scientists, procurement managers, training managers, instructors, and operational commanders. The decisions being made directly affect training and operational systems effectiveness. The right kind of human performance data can be used to guide decisions about doctrine, strategy, tactics, crew-machine function allocation, control-display design, personnel selection, training, training device and curriculum design, skill maintenance, and operational readiness. Although there are some isolated exceptions, there is a general lack of valid human performance information to help guide these decisions.

The need for improved aircrew performance measurement has been known for a long time (cf. Smode and Meyer, 1966; Department of Defense, 1968); many existing training and operational performance measurement practices do not provide the kind or quality of data required by current technology for good decision making. The need continues to be more pressing because current budgetary and fuel realities demand unprecedented improvement in the efficiency of training, skill maintenance, and operations. In order to increase efficiency and maintain (or improve) current system effectiveness, we must improve measurement. Responsiveness to this need can be seen in the ongoing performance measurement programs in the Navy, Air Force, and Army.

Measurement related to human performance in systems may be arranged arbitrarily into six classifications: (1) basic human abilities, (2) subject matter knowledge, (3) work history and experience, (4) performance effectiveness, (5) overall system effectiveness (of which transfer-of-training is considered herein as

a special case), and (6) cost effectiveness. The greatest need in aircrew performance assessment appears to be the development of valid performance criteria for training and operational environments. Thus, this discussion focuses on crew/system performance effectiveness measurement.

Aircrew personnel are system and subsystem operators who monitor, make decisions, and control during the execution of training and operational flight missions. Their performance is deeply embedded in many systems. Extracting information about crew performance from these systems requires direct and indirect measurement of many variables and poses methodological challenges. Several methods to approach the problem appear to work with various degrees of success. Substantial progress has been made in the past 4 years; we have some answers, but the evidence is not all in. There are issues that have not been well resolved, primarily due to a lack of sufficient data. The present status and future research needs of aircrew performance meausrement will be described in the following three major sections of this paper: (1) the aircrew environment, (2) an approach to measurement development, and (3) future research needs.

THE AIRCREW ENVIRONMENT

Although great technological advances have been made in aviation and airborne weapon systems, development of objective, numerical standards of crew performance has not kept pace with the technology. The reason is that there are many, many system and human variables and too many unknowns. The magnitude of the technical challenge can be introducted in the following three oversimplified views of the aircrew world: (1) the training process, (2) the aircraft and weapon system environment, and (3) the operational environment.

The Training Process

A global view of the training process is shown in Figure 1. Early in the process there is a lot of measurement, great institutional memory of data, and relatively well developed criteria such as test scores and success in school. As one proceeds toward the execution of real world missions, there is less measurement and less institutional memory, and performance criteria become more complex.

Criteria at various stages in the process may or may not be related to real world mission success. Issues such as selection based on job samples, selection for differential assignment, and attempts to identify the characteristics of crew members who have been outstanding in combat tend to reflect uneasiness that those very characteristics that would make a person outstanding in the real world mission might cause him or her to fail educational, selection, and school criteria as they exist today.

The point becomes obvious: our purpose is to predict real world aircrew performance effectiveness in order to make appropriate decisions. The further from the real world our data are taken, the less valid will be our measures, criteria, and decisions relative to that world unless predictive validities have been established. In order to establish predictive validity and to improve criteria along the way, we must find ways to measure as close to the real world as possible.

The Aircraft and Weapon System Environment

Technology itself is making the job of assessing aircrew performance and establishing criteria more complex. In response to operational needs, the technology is spurred to develop new systems that have improved performance characteristics. Compared to the past, aircraft and weapon systems are more sophisticated in every sense. So is the entire command and control system. Improved performance is provided by increased technical complexity of systems, both external to and onboard the aircraft.

One attendant price of this technology is a dramatic increase in the range and complexity of aircrew tasks in both commercial and military flight environments. Flight crews (1) must learn to use many systems in various modes, (2) must recognize system degradation and malfunction, (3) must know what can and cannot be done when subsystems or combinations of subsystems fail, and (4) must be able to take over automatic functions with operational proficiency at various semi-automatic or manual levels of control. Obviously, a great deal of system knowledge is required. There are, for example, 150 different system failures presently programmed in the F-14 Operational Flight Trainer, not including weapon system or NFO (rear seat) functions.

The environment has changed markedly from the days when aircrew performance could be judged solely on the ability to start the engines, to takeoff and land safely, to fly by instruments, to navigate, and, in combat, to drop bombs or out-maneuver a single opponent in a dogfight. These tasks remain important today, but they represent only a part of the job. Flight crews have become sophisticated electronic systems managers and analysts. To even qualify for selection in the electronic warfare area, a candidate must have both a degree in engineering and other outstanding credentials. Crew members have to recognize complex system parameters and envelopes while driving toward maneuvering adversaries. The adversary weapon system capability is also complex. The dynamic characteristics of various threats have to be recognized if the aircrew is to survive and promote overall system effectiveness.

Aircrew knowledge of the capabilities of the systems at their command and the demonstrated application of these capabilities against the performance characteristics of the adversary under a host of dynamic situations is the proper basis for aircrew performance measurement criteria. Academic knowledge alone provides incomplete assessment criteria because there are many conflicting factors and stressors in the real-time flight environment that can prevent effective recall or use of academic knowledge.

The Operational Environment

Our third view of the aircrew world focuses on operations, where the products of training, aircraft, and weapon system development come together for an operational purpose. A partial, closed-loop feedback diagram (Figure 2) is used to show that aircrew performance is embedded in a large, dynamic system composed of subsystems which are both external and internal to the aircraft.

Starting with the external subsystems, doctrine, strategy, and tactics will influence a particular mission plan. If the mission is a combat mission, then adversary doctrine, strategy, tactics, and force structures enter into consideration.

The aircrew performance of the mission will be influenced (as applicable) by weather, obstacles, the availability of friendly weapons, other aircraft in the formation, and commands issued by ground or airborne control centers. If the crew is attacking a target, then target movement and adversary weapons become strong forcing functions. Crew/system performance will be relative to those targets, taking threats and own weapon system capabilities into account.

In the aircraft, the flight crew is embedded within the subsystem display and control loop, drawing information from the visual world, radio communications, crew intercommunications, instrument displays, maps, checklists, and their experience and knowledge. The crew acquires information, processes it, and makes decisions and control inputs in accordance with generally prescribed duties and functions (represented by the crew function blocks in Figure 2). Although individual crew members are often dedicated to specific functions and subsystems, their tasks can vary with the situation. Crew control inputs are processed by the various subsystems, including the airframe. Airframe and subsystem states feed back to displays and external subsystems.

Appropriate aircrew behavior is a function of dynamically changing states of the aircraft, its subsystems, crew interaction, and a host of external commands and forcing functions. Thus, totally comprehensive performance criteria must be guided by all system states which are relevant at the time performance is measured. Literally, several hundred (thousand?) variables may be involved (Table 1); many variables are difficult or costly to measure.

As a result, quantitative aircrew performance criteria have been defined or legislated for relatively few operational tasks (such as nominal instrument flight, carrier landing, and air-ground weapons delivery) and often under highly constrained task "setup" situations such as the gunnery range. Even where criteria do exist, it is sometimes found that the measured performance of experts does not necessarily correspond with published criteria (Knoop and Welde, 1973; Vreuls, Wooldridge, Obermayer, Johnson, Goldstein, and Norman, 1976). Therefore, any assumptions made about performance criteria should be treated as assumptions unless they are validated empirically by measuring what skilled crews really do in a particular environment.

Summary

The aircrew environment from a measurement viewpoint may be characterized as follows:

1. There is more measurement and institutional memory early in training than later in training or operations.

2. Increased aircraft and weapon system sophistication has added new cognitive dimensions to flight crew tasks and workload.

3. There are many dynamically changing variables both inside and outside the aircraft that influence performance.

4. Performance criteria that may be derived from publications or subject matter experts have to be treated as reasonable and relevant, but as unvalidated in the absence of firm data.

5. The closer to the real world environment criterion measurement is taken, the better will be predictions of that world.

6. Firm performance criteria are more difficult to specify as the performance of interest approaches the operational job.

Yet, the whole training and operational system should be geared toward performance criteria which are derived from real world operations. Any approach to aircrew performance assessment must recognize and deal with these factors.

AN APPROACH TO MEASUREMENT DEVELOPMENT

We have taken the position that the end product of performance measurement development must be information capable of guiding many different kinds of decisions. In order to do this, measurement should have demonstrated diagnostic power and validity for each purpose. Diagnostic power is the ability of the measure set to describe why a performance is good or bad; it is needed most when marginal or substandard performance emerges in order to be able to prescribe solution.

It is unlikely that a single performance score can provide diagnostic information, although single scores are useful for summary information and certain higher-level decisions. Multivariate (statistical and system modelling) techniques offer the only methods that are powerful enough to capture the complexity of the real world in sufficient depth for diagnosis and yet provide single metrics (which represent performance functions composed of several variables) for higher level summary. The measurement research to date has had to improve and tailor multivariate techniques as well as develop overall methodology and, in some cases, also provide usable measures for the training and operational world.

Most performance measurement investigators understand that, ultimately, fully diagnostic measurement will have to include selected measures of (1) basic abilities, (2) subject matter knowledge, (3) past performance, work or training history, and (4) current task performance. Current task performance may be further subdivided into measure sets for airborne and simulator use because practical measurement in each environment differs, yet these measures should be related by a common subset. This discussion concentrates on current task performance with full knowledge that, eventually, all measurement domains will have to be brought together to service the need for comprehensive diagnosis.

It is just not practical to measure all relevant variables for all tasks. Measuring "everything that moves" is neither cost-effective nor necessary in most cases. Empirical measurement work (cf. Waag, Eddowes, Fuller, and Fuller, 1975; Sanders, Kimball, Frezell, and Hoffman, 1975; Vreuls et al., 1976; Lees, Kimball, Hoffman, and Stone, 1976) suggests that a greater share of the performance variance can be described by fewer than 15 measures in many flight tasks.

The measurement development method must devise a way to sample those tasks and measures that are most important for describing, understanding, and predicting crew/system performance. When a sampling approach is taken, there are many assumptions made in the process, and it must be demonstrated that the sample is valid. Establishing the validity of performance measurement samples is a challenging issue that is well described by Waag and Knoop (1977).

The best method to derive the measurement sample has also been an issue of concern. There seem to be two major alternatives: (1) measure "everything that moves" initially and devise computer algorithms to decide what is important, or (2) initially reduce the world of all possible measures to a smaller set of measure candidates (by some means other than empirical data collection) and test the resulting measure candidates empirically to establish the final, useful measures and format. Some of the pros and cons of these two alternatives are discussed by Waag and Knoop (1977). Suffice it to say here that the first approach, although scientifically appealing, requires enormous amounts of initial data collection. The second approach is more practical at this time.

The recommended approach for development of performance measurement can be thought of as containing five steps: (1) measurement analysis to define a reasonable, initial set of possible measures and standards for describing performance using mission and task analytic methods, (2) design and development of the data acquisition system (automatic or manual), (3) collecting actual performance data, (4) using statistical analyses to select those measures and interrelationships between measures which are most important for describing and diagnosing performance, and establishing various forms of empirical validity, and (5) testing the resulting measurement for utility. These methodological steps are discussed in the following subsections.

Measurement Analysis

Measurement analysis must decide what to measure and how to measure it. The keys to good flight performance measurement include: (1) adequate sampling of decisional, procedural, mission-related, and perceptual-motor skills as common as possible to a range of tasks, (2) unambiguous definition of when the observations are to be taken, (3) explicit definition of the indices of desired performance, (4) comparision of actual to desired performance at the fewest critical points that will yield reliable scores, and (5) due consideration of the possibility for different information formats, taking into account the information needs, capacities, and limitations of the user.

The first area of concern for measurement analysis requires identification of objective behavioral standards against which resulting performance may be compared. Extensive interaction between the measurement analyst and one or more subject matter experts is required by the analytic process. The status of subsystem variables external to the aircraft are defined through mission and mission segment analyses. Aircrew tasks are then defined through various forms of task analysis relative to those missions and segments. Some mission segments and tasks of interest to aircrew performance measurement are contained in Table 2.

In many flight regimes the standards are well enough understood that an Instructional System Development (ISD) type of process can result in a profitable definition of measurement (cf. Baum, Smith, and Goebel, 1973; Northrop, 1976). A mission-maneuver-measurement analysis framework (Obermayer, Vreuls, Muckler, Conway, and Fitzgerald, 1972) can shorten the formal ISD process considerably by proceeding more directly to measurement specification without formal documentation of intervening behavioral objectives. Whatever the process, it is helpful to have a well developed framework for measurement at the onset of the analysis.

A measurement framework has been established (Vreuls, Obermayer, Goldstein, and Cauber, 1973) which relates system performance and human behavior to segments of maneuvers constituting a mission. The structure, derived from earlier work (Benenati, Hull, Korobow, and Nienaltowski, 1962; Knoop, 1968), permits the measurement of a variety of tasks and performance dimensions in order to describe unique as well as common aspects of maneuvers. It requires each measure to be defined in terms of the following five determinants:

1. Measure segment.
2. State variable or variables.
3. Sampling rate.
4. Desired value.
5. Transformation.

A measure segment is any portion of a maneuver or mission for which desired student behavior or resulting system performance is relatively constant or follows a lawful relationship from beginning to end, and for which the beginning and end can be unambiguously defined. Measure segments may overlap or they may be simple one-time events. The segment is defined by explicit measurement start/stop logic.

A state variable is any quantifiable index of (1) vehicle states in any reference plane, (2) personnel physiological states such as heart rate or eye movement, (3) control device states such as stick, pedal, or switch positions, or (4) data from a source outside of the immediate vehicle, such as external variables or even personal history variables. Multiple variables may require special mathematical or logical treatment before or after they are treated by the rest of the process. For example, one may wish to combine variables to form a functional relationship before sampling or comparing that function to a desired value.

The sampling rate is the temporal frequency at which a parameter is recorded or examined, primarily by automated measurement systems. Guidelines for selecting sampling rates are readily accessible from standard practice and consideration of human/system dynamic response characteristics. Where temporal samples are required for manual measurement (and the needs do not fit well within the definition of measure segments), observer workload and capacity must be carefully considered.

In almost all cases, a state variable has no utility unless compared with a desired value to derive an error score. Desired values may be determined analytically or empirically.

Finally, a transformation is defined as a mathematical treatment of the error score, which may be just its value or absolute value, or may include computation of out-of-tolerence conditions, measures of central tendency, variability, frequency content, departures from norms, etc. Common transformations are shown in Table 3.

This structure appears to be a good analytic/descriptive framework for computer and manual implementation of measurement. For manual measurement, all of the same principles apply; however, special attention has to be given early in analysis to the number and kinds of discriminations required of the human observer, the format and human engineering of the measurement instrument, the imposed workload, and user acceptance.

Although profitable for initial measurement definition, task analytic procedures have inherent limitations. In many cases, especially in maneuvering flight, there are wide ranges of permissible and equally successful behaviors. There are wide differences of opinion regarding appropriate behavior and standards among subject matter experts. There are dynamic performance requirements that are hard to specify with the method. Often, value judgments are required by both the subject matter expert and the measurement analyst. There are no firm criteria to guide these judgments short of experience with actual data collected from skilled and unskilled performances of the task. These data are rare indeed.

The result is some uncertainty about the relative importance of measures. The analyst either (1) accepts subject matter expert opinion where there is reasonable agreement, running a slight risk of missing something important where consensus is not possible, or (2) deliberately overmeasures in areas of uncertainty, letting the empirical portion of the process determine which of the measure candidates account for most of the performance. The safest course of action is to overmeasure slightly where there is uncertainty.

An inherent limitation of the measurement structure is defining unambiguous conditions for measure segmentation, the measurement start and stop rules. It is often very difficult to decide when a maneuver begins and ends in real-time. For example, unless you know a pilot's intentions, false turn starts are easy to record in turbulence (Knoop and Welde, 1973) and there may be many "tops" to a barrel roll attack[1] depending on pilot/aircraft stability. Analytically defined rules often fail when they are tested empirically.

The inherent limitations of task analytic procedures do not invalidate them as a reasonable starting point for measure development; however, these limitations soften the precision of the method and point to a serious need to research better methods to analytically prescribe measurement.

The amount of emphasis that should be placed on analysis for measurement is a matter of establishing the degree of utility of the analysis output, given other means to derive information such as empirical data collection. Even small amounts of real data can save hours of analysis effort. Because of the number of assumptions and uncertainties involved, our experience suggests that too much pencil and paper analysis is not cost-effective. A thorough analysis should be done, but one should move promptly toward empirical data collection.

---

[1]Spring, W. Personal communication on the Northrop/Navy ACM studies, November 1976.

Measurement System Design

Analysis for measurement should produce specifications of desired measurement. Implementing these specifications can be a straightforward or extensive effort, depending upon the mission task and environment in which measures are to be taken, the sources of measurement (manual, photographic, or automatic), and the uses of the data once they are acquired. Not only do data collection systems have to be designed, but also, systems have to exist or be produced to check, verify, smooth, correct, label, catalog, store, and extract data for a variety of analyses. A good example of the considerations for measurement systems may be found in Obermayer and Vreuls (1972) and Obermayer, Vreuls, Muckler, Conway, and Fitzgerald (1972).

Parenthetically, it has been well stated by Roscoe (in press) that performance measurement does not have to be automated to be objective, reliable, and valid. Objectivity requires that performance standards can be observed publicly, as opposed to subjectivity, which is private. Ratings will be subjective if observers are required to employ their own private standards of correct or desired performance. Two or more observers evaluating actual performance or reviewing records of performance will quite likely arrive at the same judgment of the performance quality if (1) they have done so without distraction or personal hazard, (2) they have agreed on the standards, explicit indices of desired performance and conditions for measurement, (3) it is physically possible for them to act as reliable sensors, and (4) they have been "anchored" by having observed the full range of performance.

The lack of reliability of observer, instructor, or check crew member ratings most frequently will occur because the above conditions have not or cannot be met. In some flight situations an observer does not have the span of observation (or perceptual/cognitive bandwidth) required to canvass the entire field of critical dependent variables at precisely the critical times and, simultaneously, to perform other required duties. When this is the case, automated systems become highly desirable. Automated measurement systems are a necessity for training or testing systems that use adaptive or automated techniques.

The current state of technology permits relatively straightforward development of measurement systems; however, there are some cautions. It is one thing to instrument a system or to develop scoring forms for research purposes; on a one-time basis, data may be collected by trained observers or by one-time instrumentation systems. It is another matter to develop a real-time measurement system to be used in flight training or in operations on a regular basis by non-researchers as well as researchers. We assume that measurement work must lead toward use by the operational or training community. As such, the design effort must be guided by practical realities as well as the information desired.

Developing real-time or near real-time measurement systems for the airborne environment requires special caution. If recorders, cameras, or computers are being used, there are considerations of weight, size, packaging, power, heat, vibration, noise, and airworthiness (of both the measurement system and the systems it may be tapping for data) that must be attended to and resolved. Design trade-off decisions have to be made on a number of issues, including whether the onboard system simply records raw data or produces the results of measurement. If the onboard system only records raw data, a greater data reduction and processing burden is shifted to ground systems. An example of one approach to airborne instrumentation may be found in Knoop and Welde (1973).

Modern flight simulators often provide an environment that is better suited to measurement but, unless they have been specifically designed for measurement, there are typical problems encountered. Often, one finds a lack of adequate documentation, timing problems, disagreement between the math model and measured system responses, measurement scaling and transformation irregularities, A/D or D/A conversion discontinuities, distributed processing, and limited resources to handle additional computational demands, to store, or to output desired information. Typically, these are not insurmountable problems, but they have to receive their due attention.

Developmental measurement systems should be designed to facilitate change because the analysis process that produced the specification is less than perfect. Once built, both manual and automatic systems require developmental testing to ensure their operation. Automatic systems require special attention to calibration, data identification, noise, drift, and measurement start/stop logic. Manual systems require special attention to the reliability of observations, imposed workload, training of observers, and design of scoring forms. Both systems require tests of data reduction or processing systems. Developmental testing often reveals a need for change even before formal data collection begins; the experienced investigator plans on it.

Cost and time saving short-cuts may be taken with experimental measurement systems in order to collect empirical data for measurement development. However, when the final design becomes operational, it must conform to a host of military specifications that include reliability, maintainability, and logistic requirements (spares, documentation, training of personnel, maintenance of hardware and software, etc.) which are not always recognized by researchers, but should be. Also, it is not always recognized that, today, the cost of computer software far exceeds the costs of hardware for many special-purpose applications. If, for example, modification of a flight simulator is desired for measurement, it may be an order of magnitude cheaper to strap-on special purpose mini- or microcomputers than to reprogram the existing software. Careful design trade-off decisions are required.

The problems of developing airborne or flight simulator data collection or measurement systems are not as staggering as they were just a few years ago. Advances in technology, the sophistication of systems already onboard many aircraft, the existence of instrumented air combat maneuvering ranges, and the capabilities of modern digital flight simulators have markedly increased the feasibility and practicality of special-purpose measurement systems. The costs of such systems may be insignificant compared to the costs of making decisions based on incomplete information.

Data Collection

All of the familiar rules of experimentation in the laboratory also apply to measurement data collection in operational or training environments; however, the experimenter has less control of the equipment, people, and schedule. The orchestration requires thorough planning and coordination well in advance of data collection, and requires a thorough understanding of the environment. Special attention must be given to the impact of data collection on ongoing activities. Although everyone tries to avoid it, schedule collisions are inevitable due to equipment malfunction, weather, and operational mission requirements. The seasoned investigator develops contingency plans and avoids overzealous data collection schedules.

Experimental designs for measurement development studies have been relatively straightforward. Independent variables such as pilot skill level, time in training, aircraft weight and center-of-gravity, turbulence, and workload (such as command pacing) have been deliberately manipulated. For measurement studies, it is not always practical to manipulate many independent variables; where variables cannot be controlled, they should be measured. Often, statistical control is possible during measure selection analyses.

It is now possible to use repeated measure experimental designs (where each subject contributes more than one observation) for measurement studies employing multivariate analysis techniques. Recent work by Vreuls (1976) and Wooldridge, Breaux, and Weinman (in press), based in part on earlier work of Schori and Tindall (1972), permits repeated observations on the same subject in one data collection session and again at a later time. This methodological breakthrough has reduced subject requirements by an order of magnitude.

The data collection requirements for measurement studies using multivariate analysis techniques may be slightly greater than other kinds of research because of the number of variables involved. The driving requirement will be the product of (1) the number of initial measures, (2) the number of independent variables, and (3) a replication factor. Lane (1971) has shown that, to avoid shrinkage and overfit, there should be seven to nine times as many observations as initial measures; thus, 20 initial measures might require 140-180 observations on the same task. However, Wooldridge et al. (in press) examined the problem using data obtained from 12 pilots training on four instrument flight maneuvers in an F-4 simulator. Given our measure selection techniques (described later), shrinkage calculations have suggested that only three to five times as many observations as initial measures would be required, depending upon how well the known variables account for performance variability. Thus, we should be able to reduce the Lane criterion for observations by one-half.

Further reductions in the number of required observations may be possible if, during initial data collection or screening prestudies, it is found that less than the full set of measures accounts for most of the performance variance. In order to hold data collection requirements to the absolute minimum, screening methods must be employed to iteratively collect data and perform analyses, repeating the process only as it is necessary to do so. The use of iterative screening methods, however, can create scheduling uncertainties that are not practical in operational environments. Often the best that can be done is to perform successive and simultaneous analyses as the data are collected, if it is possible to do so.

Two final comments on data collection for measurement studies cannot be overemphasized. First, if a variable is not controlled, measure it. We strongly suspect that certain variables such as subject experience, age, recency, motivation, time of day, and weather make a difference in performance. Measuring them may improve substantially the amount of variance accounted for. Multivariate analysis techniques permit limited statistical control of the otherwise uncontrolled experiment. Secondly, although strictly mechanical, a well planned data cleanup and editing process, as close as possible to the time of data collection, can save hours of agony later on. A good data edit and management system is quite necessary with the volume of data that measurement studies process.

Selecting and Weighting Measures

The end goals of the measure selection process are (1) to validate the assumptions that have been made in the task and measure definition sampling process, (2) to find the smallest comprehensive subset of candidate measures that will adequately describe performance on a given task (to reduce observation and output requirements) and (3) to "classify" and weight measures for a specific purpose.

Crew/system performance is multidimensional. The importance of any particular measure with respect to performance diagnosis usually cannot be stated without consideration of all other measures of the performance and situational set. Measurement analysis must find these relationships and express them in a coherent format that (1) explains to the measurement specialist what is happening in performance space, and (2) can be used as a basis to derive usable measures and formats for the operational world.

Measure Selection Criteria. Several criteria for selecting and weighting measures have been used in various combinations. The more common criteria include (1) the amount of variance accounted for by each measure, given all measures of the set, (2) correlation with instructor or observer scores, (3) terminal performance (such as carrier landing or ACM engagement outcome), and (4) the ability of the set to discriminate between performance by various skill levels (e.g., student versus expert; early versus later in training). There is a need for better criteria to establish validity of measures for many purposes (Waag and Knoop, 1977); however, a great deal of work can be accomplished with existing criteria.

Measure Selection Methods. Work to date suggests that the measure selection process may be conducted in a variety of ways, as long as empirical validation is performed. Slightly different multivariate approaches to measurement development can be seen in the work of the following investigators:

Locke, Zavala, and Fleishman (1965) used factor analytic techniques to relate pilot helicopter performance to task and maneuvering factors. They verified that measurement of a given task in the context of one maneuver was correlated strongly to performance of the same task in another maneuver. They suggested that basic human abilities could be mapped into task factors (see also Fleishman, 1967).

Connelly, Schuler, and Knoop (1969) derived a set of adaptive mathematical models which organized measured pilot performance in a simulator based on simultaneous scoring by several instructors. This study represents the approach of collecting many of all possible measures and using mathematical algorithms to decide what is important.

Brictson, Burger, and Wolfeck (1973) used multiple regression techniques and instrumented and manual measurement to develop final manual scoring of carrier landing terminal performance. The importance of this work is that it carried the measurement development and validation process to its logical conclusion by producing the Landing Performance Score which is currently employed by the fleet.

Knoop and Welde (1973) used regression techniques and functional relationships between variables to derive measures for two inflight maneuvers, the lazy eight and barrel roll. Their work encountered and solved several difficulties of inflight measurement and demonstrated the feasibility of instrumented inflight measurement in the aircrew training environment.

Waag et al. (1975) employed multiple regression and correlation techniques and automated and manual measurement. They used correlation with instructor scores and the ability of the measure set to discriminate pilot skill levels as criteria for selecting measures for several instrument flying tasks in a simulator. The work has been continued to include more flight tasks.

In an ambitious attempt, Carter (Northrop, 1976) has 552 automated measures and 100 instructor scores of 208 barrel roll attacks against an autopilot driven bogey in a simulator. His measure selection procedures included multiple regression, step-wise multiple discriminant, "jackknife" partial cross validation, and ridge regression analyses of selected data subsets.

Sanders et al. (1975) and Lees et al. (1976) used manual and airborne instrumentation measurement and step-wise discriminant and multiple discriminant techniques to select measures for low level and Nap-of-the-Earth (NOE), day and night helicopter flight. The final discriminating measure sets include a substantial portion of control input measures.

Vreuls et al. (1976) and Wooldridge et al (in press) have used highly tailored multiple discriminant and multiple regression techniques to select measures for four instrument flight tasks in an automated flight simulator, based on the ability of the set to discriminate skill level. A partial validation of the technique demonstrated its utility for automated training of civilian private pilots. Method validation using military pilots is continuing with measure selection studies for automated simulator GCA training at Luke Air Force Base.

Several improvements to multivariate measure selection methods have been made. Our multiple discriminant and multiple regression programs have been tailored (1) to remove highly correlated measures, (2) to eliminate performance outliers that dangerously distort a least-squares fit, (3) to correct for repeated observations (discussed earlier), (4) to improve orthogonality and predictive validity by stabilizing weighting coefficients using ridge regression (Hoerl and Kennard, 1970) and (5) to develop a procedure to automatically remove measures based on their commonalities.

General Results. There are some general results of the above studies that are of interest to measurement development. First, control input measures are often important contributors to discriminating the differences between skilled and unskilled performance. Secondly, performance differences on tasks often are revealed in unexpected ways; for example, during a level, 30 degree bank turn, major portions of the performance differences may be found in altitude error, not bank error, but this kind of result is airframe and task specific. Thirdly, variations in the task--such as (1) making a 60 degree bank turn instead of a 30 degree bank turn, (2) the addition of turbulence, (3) shifting the center-of-gravity, or (4) varying workload--change the nature of the task; major changes occur in the discriminating measures sets and the measure weighting coefficients. Finally, the proper weighting of measures for combining them into

a single score cannot be predicted analytically; frequently, criteria such as "5 degrees of heading error is equivalent to 200 feet of altitude error" are not supported by empirical selection results. Multivariate methods are able to address these kinds of problems and produce usable measurement.

However desirable multiple discriminant or multiple regression methods may be for measure selection, there may be times when it is not practical to collect a sufficient amount of data to use them. In these cases, measures may be expressed in terms of their departures from the norm as Z-scores, and individual Z-scores may be averaged to produce a usable, single score metric. This procedure has been shown (Vreuls et al., 1976) to be a good fall-back measurement selection method; it produces measures that are superior to those derived by analysis alone but less satisfactory than those derived by the full multiple regression or multiple discriminant methods.

We have observed that statistical significance of a particular measure is an insufficient criterion for measure selection. Statistical significance, by itself, can be produced by ponderous and costly replication if necessary, with the possible result that a statistically significant measure may account for an infinitesimal amount of variance in the performance space. It is more important for measures to describe well the performance space (i.e., account for variance); if they do so, significance usually will follow, but the converse is not always true.

Measure Selection Method Improvement. Multivariate measure selection methods may be improved further, but some research is needed in at least two areas. The first area has to do with unit weighting (equal weighting of all predictor variables). Flight data deal with measures of different scale characteristics and perhaps different amounts of importance for specific purposes. Yet, it is useful to combine measures into a single score. Current methods provide weighting coefficients based on least-squares regression criteria for this purpose. Given a reasonable number of measures in a set, there is evidence that unit weighting may lead to better prediction than weights based on least-squares criteria (Einham and Hogarth, 1975; Wainer, 1976). However, if one does not use weighting coefficients, desirable features of scaling measures and weighting their importance are lost. Ridge regression may circumvent the problem, but we don't know that to be true and intend to look into it.

The second area of research has to do with the best way to relate dependent and independent variables from several sources for measure selection purposes. Given that certain independent variables (such as flight time, aircraft center-of-gravity, turbulence, and, possibly, academic grades and abilities) influence performance, how does one best account for these variables in measure selection? One can simply include them as dummy variables in a regression equation along with measures. However, it may be better to use canonical correlation techniques, placing the independent variables on one side of the canonical equation and dependent variables on the other. Development work is needed to insert data conditioning, measure elimination, and ridge regression techniques into the canonical method and to validate the procedure. Further thought in this area may lead to better diagnosis techniques.

Toward Greater Generalization. It may be important to distinguish the difference between the best possible measure set that results from initial measure selection analysis and the finally implemented subset for a particular application.

The best possible measure set is a sample that has been produced by task analytic procedures and empirical measure selection, conditioned by what was possible to measure for research purposes. There are differences between measures that can be taken during research and practical measures for operational use.

Where such differences exist, searching for the best possible set improves the generalization of data and permits quantitative relationships to be established between that which is best and that which is practical. Given quantitative relationships, it may be possible to translate data (1) between the two domains, (2) between flight and flight simulators that may have slightly different measure subsets, and (3) perhaps between operational and research environments for a variety of other purposes. Also, it gives the investigator a better feeling for the capabilities, limitations, and validity of the operational set that emerges from final testing.

Product Testing

To be complete, measurement development should include a final phase of testing. Two different kinds of tests appear desirable: developmental tests and system effectiveness tests. Each is briefly described below.

Developmental testing may be integrated with earlier steps in the measure development process, may be done after measure selection analysis, or both. During developmental testing, the products are exposed to the user in their final form but with an eye toward change and improvement. Although the user is involved in major stages of development, it is often the case that the impact of the product cannot be appreciated fully until it is put to use. Once put to use, the user often will suggest major improvements in the format, order, priority, and use of information. A good measurement program should plan for a reasonable period of product improvement testing.

System effectiveness tests can reveal what is gained by measurement development research. For example, it would be helpful to know the effect of different measurement approaches on resulting changes in performance quality, training time, transfer-of-training, personnel utilization, device or aircraft utilization, cost of operations, or similar metrics. These kind of data provide researchers and managers with a better idea of how to allocate always insufficient resources of money, facilities, personnel, and time. System effectiveness tests can be viewed as a useful form of empirical validation.

System effectiveness tests can be conducted concurrently or longitudinally as long as sufficient data are available to make comparisons and to statistically control for confounding effects. One simply measures system effectiveness (and any confounding variables) before and after the results of measurement development research are installed in operation. If cost data are available, then cost-effectiveness may be calculated also.

One system effectiveness test showed that empirical measurement development caused a 40 percent reduction in the time-to-train (the same performance criteria in automated flight simulator training) compared to the initial measurement which was derived by analytic means alone (Vreuls et al., 1976). This is the only study we know of that has shown an improvement in training effectiveness as a direct result of improved measurement. Obviously, we would like to see many more.

Summary

One approach to measure development has been used to describe what is known or not known about aircrew performance measurement along with the many considerations that enter into the job. Some of the key points may be summarized as follows:

1. Measure Definition--Existing techniques to sample tasks and measures do work, but the results must be validated empirically.

2. Measurement Systems--The technology is markedly improved, making better instrumentation possible in aircraft and flight simulators. Manual measurement is also possible, but stringent criteria must be met to ensure objectivity and reliability. Systems to check, process, and extract information must be provided.

3. Data Collection--The number of observations required for measure selection is slightly greater than requirements for other kinds of research. Measurement in operational environments requires attention to practical factors.

4. Measure Selection--Methods are well established, but there is room for improvement. Especially needed is more work on criteria for measure selection for a variety of information purposes and for better performance diagnosis.

5. Product Testing--Developmental tests are important. System effectiveness tests are a useful form of empirical validation.

Considerable progress has been demonstrated over the past several years, but there is much more to do.

FUTURE RESEARCH NEEDS

There is sufficient evidence of all kinds to know that performance measurement is at the very core of all kinds of research, procurement, training, and operational decisions. Improvements to measurement can and have been made. When measurement is improved, there is also improved system effectiveness.

Quantitative performance effectiveness data on a task, mission, and aircraft basis are sparse. This makes it difficult to supply validated performance criteria to assist research on the following kinds of training questions: Should expensive motion and visual systems be purchased for flight simulators? If so, what components or features are really needed? What aspects of simulator fidelity affect transfer-of-training? Should we augment flight simulators with cues that are not present in the real world? Which of these features are necessary for skill maintenance? Should tasks or whole missions be taught in flight simulators? What is the best mix of simulator and flight time for initial training and for skill maintenance? What is the best mix (in terms of the amount of time, subject matter content, and sequencing) for academics, part task trainers, simulators, aircraft, and other media for both training and skill maintenance? Each of these questions should be addressed on a task, mission, and aircraft basis.

Data are also needed to address questions more directly related to measurement issues, such as the following: What are the important components of operational proficiency? What distinguishes good and bad operational performance? How is such performance related to operational readiness? How do we best characterize

operational performance as criteria for various research purposes? What are the cost-effective performance levels for each stage of pilot or NFO training? What are quantitative fleet expectations for incoming replacement crews? How do they compare with experienced crews? What do performance levels and criteria throughout training, research, and operational environments have to do with real-world mission success? Again, each of these questions should be addressed on a task, mission, and aircraft basis.

In order to provide quantitative data to address these kinds of questions, aircrew performance measurement research needs (1) empirical data, (2) better analytic methods, (3) measurement standardization, and (4) personnel.

Empirical Data

Crew/system performance data are needed to improve measurement and methods. Data are needed from many training and operational sources such as personnel records, selection batteries, academics, part task trainers, flight simulators, aircraft, controllers, air combat and air-ground ranges, and operational exercises. Data are needed on a task, aircraft, and mission basis to explore and build better criteria for decisions, better analytic methods, and better performance diagnosis techniques. Some data have been collected, but much more are needed.

Better Analytic Methods

Possible improvement to existing multivariate measure selection and weighting methods in two areas have been discussed. In addition, work is needed to relate performance measurement to system effectiveness and other criteria. For example, it would be useful to know which performance dimensions are most important to minimizing time-to-train criteria. Especially important might be the relationships between basic abilities, subject matter knowledge, work experience, performance effectiveness, and system effectiveness. Quantitative multivariate maps between these domains have yet to be established. If it is at all possible to do so, they might provide insight into the problems of manipulating particular variables in the training or skill maintenance process in order to maximize system effectiveness criteria. It would be a challenging line of inquiry.

Just as challenging is the need to develop better analytical methods to reduce our dependence on brute-force empirical data collection. Existing task analytic methods for measure definition tend to be more of an art than a science, primarily because there are no guidelines for the value judgments that are required. The method tells us what we might want to measure, but it lacks precision and does not tell us what we can ignore and still account for most of performance space. There is room for invention.

One way to approach the problem would be to provide better information and guidelines to the measurement analyst. The information the measurement analyst needs is frequently not contained in published reports. Furthermore, a format that is specific to the needs of the measurement analyst is required. Better measure definition analysis might result if, in one source, there was a collection and classification of empirical data and measure segmentation logic across the spectrum of tasks, missions, and aircraft in training and operations. Building, publishing, and periodically updating an aircrew performance measurement data base and handbook could provide analysts with quantitative guidelines to replace current value judgments. In addition, the data base might be used for special analyses as required. The job would be a big one, and a central ingredient to make it work would be the standardization of measurement.

Measurement Standardization

There is a serious need to partially standardize measurement and conditions under which measures are taken across all three services. It is often difficult to compare data across studies because the same measures and measure segmentation rules are not used. A plea for total standardization of measurement would be naive because every research effort and operational environment has its own particular information needs. However, partial standardization of common measure subsets may be relatively straightforward and useful and may be approached modestly over time.

There are today common tasks, mission segments, measure segments, and certain measures that a group of experienced measurement analysts and subject matter experts would agree are essential. Standards for these measures could be defined. Over time, the "standards committee" could meet to review the state of knowledge and to add further definition. The key to making it work would be to define only the desired, common subset; each investigator and particular application would be free to include other measures that were deemed important. Also, it would be helpful if each investigator would report the values of independent variables that were held constant. The process may be viewed as no different than standarization efforts in the physical sciences and engineering; it deserves serious consideration.

Personnel

People do research. The aircrew performance measurement area appears to require a special breed. The job requires competence in experimental methods of psychology and engineering, multivariate statistical and system modeling methods, and task and mission analytic methods. It requires a working familiarity with physical and electronic system hardware, computer software and programming methods, flight, and the aircrew training and operational environment. Seasoned investigators have competence in many areas; but seasoned investigators are few in number and, although the list is growing, many more are going to be needed.

A Final Comment

The whole training and operational system does work today. Aircrews are trained and operations performed with reasonable levels of system effectiveness. This is because a myriad of dedicated professional people have a lot of personal knowledge and use it to make the system work. What performance measurement is trying to do is make that knowledge more public and quantitative in order to provide information that will improve efficiency while either maintaining current levels of system effectiveness or improving them.

When measured, it is sometimes found that performance is different from commonly held views. Everyone in the operational, training, research, and management community should understand that what people really do might be different on occasion from the "school solution." Often, performance is much better than expected; sometimes it is different, but usually for good reasons. For example, performance may exceed common error criteria for instrument flight because when a radar controller calls out, "Fast moving traffic, twelve o'clock, one mile, opposite direction, altitude unknown," there is only one criterion.

Table 1

Partial List of Variables Implied by Figure 1

OWN AIRCRAFT (INTERNAL) SUBSYSTEMS	
Pitch and Roll Attitude	
Angle of Attack	
Heading/Turn Rate	
Altitude/Altitude Rate	
Speed/G-Forces	
Configuration	Gear/flaps/spoilers/speed brakes/wing sweep/hook External/internal load, weight, c.g.
Communication/Data Link/ Transponder	Frequencies/mode and code External and intercom message content and load
Weapon System Status (including associated radars, IRs, ECM, look envelopes, tuning)	Operating modes Firing sequences, deployment status, firing envelopes Received signal characteristics, system control, inputs, sequences
Fuel State	Fuel flow Fuel remaining Transfer status/external stores
Control Inputs	Pitch/roll/pedals/thrust Weapon firing All subsystem controls (switch positions, etc.)
Navigation Systems States	Tacan Vor OBS/frequency DME DISTANCE ADF frequency/bearing ILS localizer/glideslope/frequency Inertial system Lat/Lon-update actions Other-LORAN, Celestial, etc.
Other Subsystem States	Hydraulic, status, pressure Electrical, status, load Display systems, mode, status (HUD and electric displays) Thrust, egt, fuel flow (tit), rpm, outlet port status Autopilot and stability augmentation system-modes Proximity warning Life support systems-pressurization, oxygen, temperature

Table 1

Partial List of Variables Implied by Figure 1 (Cont.)

EXTERNAL SUBSYSTEMS	
Weather	Ceiling/visibility/wind/turbulence Cloud height/layers/icing/temperature/density Severe weather areas Seastate
Obstacles	Terrain Man made objects
Ground/Air/Sea Controllers	Traffic advisories Flight/mission plan modification/clearances Heading/altitude/altitude rate/speed commands Target states Track/altitude/speed/number Maneuvering/jamming Probable actions
Carrier	Pitch/roll/heave Track/speed/latitude/longitude
Formation Aircraft	Heading/speed/altitude (if joining) Relative/desired position own aircraft Intentions
Target	Speed/altitude/track/number/type Probable actions/weapon/performance Relative position/energy level/fuel state Other adversary weapons system status/envelopes
Mission Plan	Flight profile/course(s)/altitudes/speeds Checkpoints/rendezvous/refueling/IP/target/marshall Winds/weather Related force structure operations Communications

Table 2

Some Mission Segments/Tasks of Interest

Pre-Takeoff	Mission Planning System Checks, Alignments, Clearance
Takeoff or Launch	Liftoff and Climb Profile Aircraft Configuration Changes
Navigation	Normal Low Level or NOE Map Interpretation
Formation Flight	
Refueling	
Surveillance	
ECM, ECCM	
Air Drop	
Combat	Air-to-Air Air-to-Ground/Sea/Undersea
Recovery, Approach, and Landing	Onshore Carrier
Other Tasks Applicable Above	Instrument Flight Communications Crew-Coordination Susbystem Operations and Procedures Normal Emergency Degraded Systems

Table 3

Common Measure Transforms

Time History Measures	Time on target
	Time out of tolerance
	Percent time in tolerance
	Maximum value out of tolerance
	Response time, rise time, overshoot
	Frequency domain approximations
	  Count of tolerance band crossing
	  Zero or average value crossings
	  Derivative sign reversals
	  Damping ratio

Amplitude-Distribution Measures	Mean, median, mode Standard deviation, variance, range Minimum/maximum value Root-mean-squared error Absolute average error

Frequency Domain Measures	Autocorrelation function
	Power spectral density function
	  Bandwidth
	  Peak Power
	  Low/high frequency power
	Bode plots, Fourier coefficients
	  Amplitude ratio
	  Phase shift
	Transfer function model parameters
	  Quasi-linear describing function
	  Cross-over model

Binary, Yes/No Measures	Switch activation sequences Segmentation sequences Procedural/decisional sequences

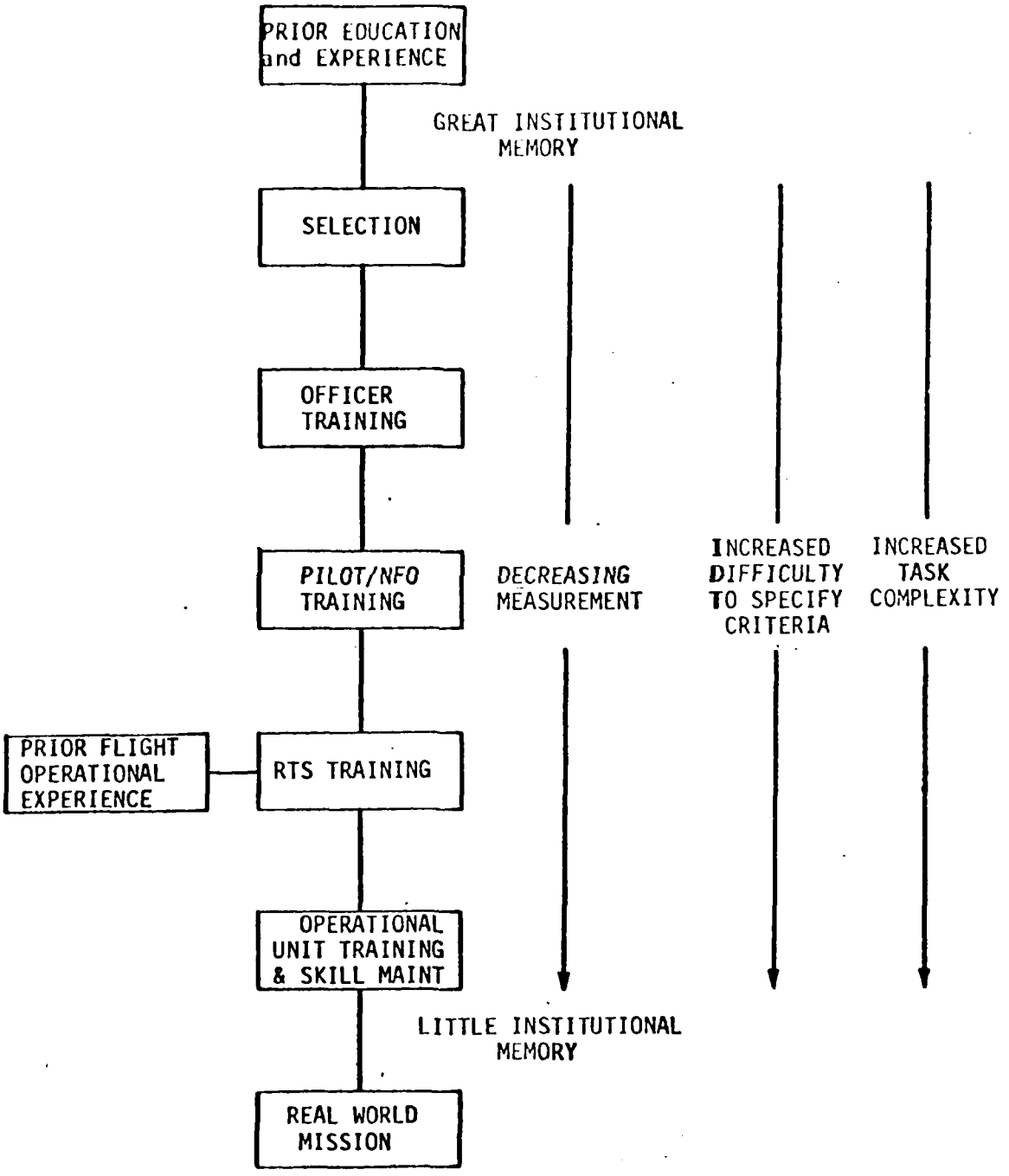


Figure 1. The Training Process

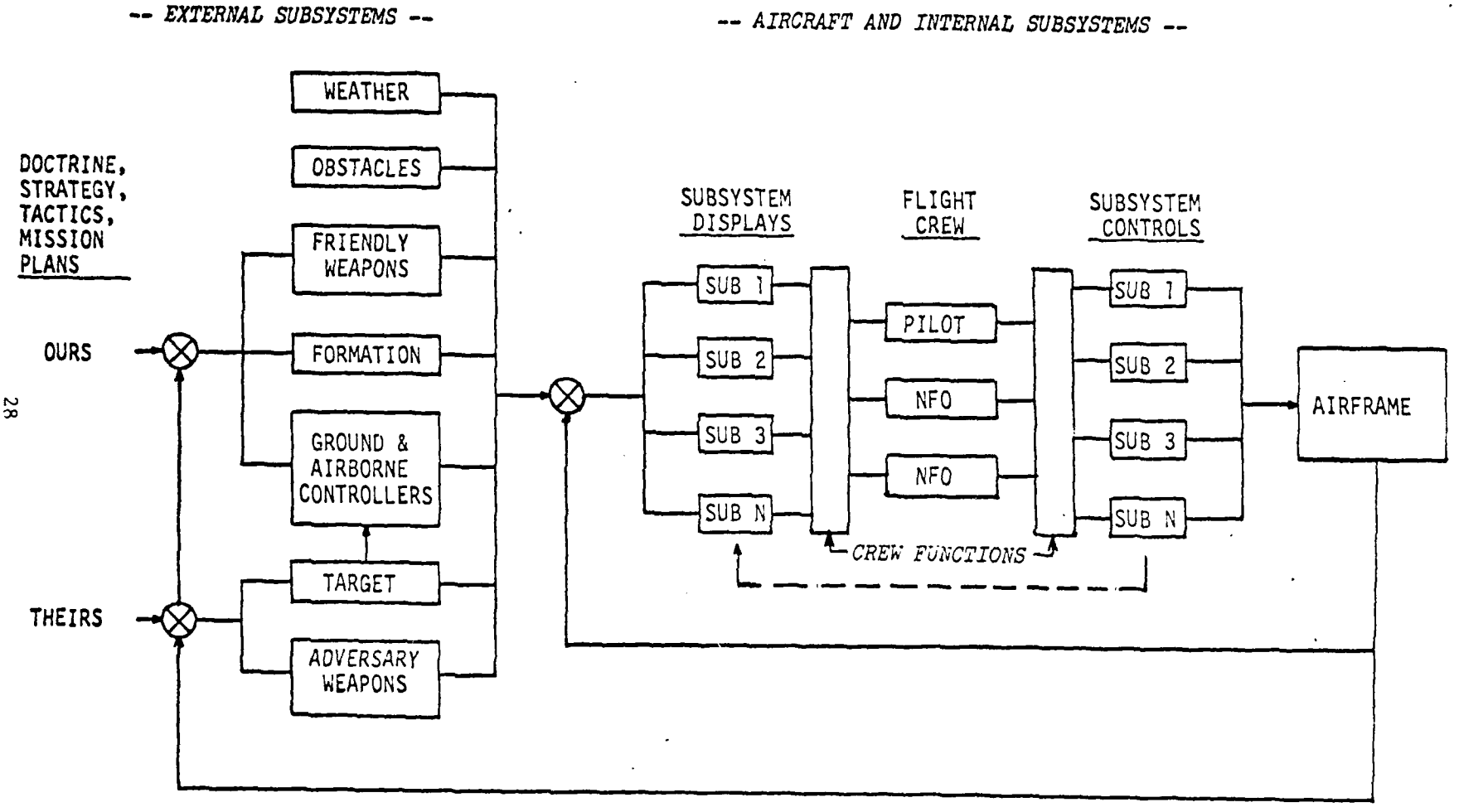


Figure 2. A Partial Closed-Loop Feedback View of the Crew Position

REFERENCES

Baum, D. R., Smith, J. F., and Goebel, R. A. Selection and analysis of UPT maneuvers for automated proficiency measurement development. USAF: AFHRL-TR-72-62, July 1973.

Benenati, A. T., Hull, R., Korobow, N., and Nienaltowski, W. Development of an automatic monitoring system for flight simulators. USAF: MRL-TDR-62-47, May 1962.

Brictson, C. A., Burger, W. J., and Wulfeck, J. Validation and application of a carrier landing performance score: The LPS. Dunlap and Associates, La Jolla, CA. March 1973.

Connelly, E. M., Schuler, A. R., and Knoop, P. A. Study of adaptive mathematical models for deriving automated pilot performance measurement techniques. AFHRL-TR-69-7, Vol. I and II, Air Force Human Resources Laboratory, Wright-Patterson AFB, Ohio, 1969.

Department of Defense. Review of tactical jet operational readiness training (U). Office of the Secretary of Defense, Washington, November 1968.

Einhorn, H. J. and Hogarth, R. M. Unit weighting schemes for decision making. Organizational Behavior and Human Performance. Vol. 13, 1975, pp 171-192.

Fleishman, E. A. Performance assessment based on an empirically derived task taxonomy. Human Factors, Vol. 9, No. 4, August 1967.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Applications to nonorthogonal problems. Technometrics, Vol. 12, No. 1, February 1970, pp 69-82.

Knoop, P. A. Development and evaluation of a digital computer program for automatic human performance monitoring in flight simulator training. USAF: AMRL-TR-67-97, August 1968.

Knoop, P. A. and Welde, W. E. Automated pilot performance assessment in the T-37: A feasibility study. AFHRL-TR-72-6, Air Force Human Resources Laboratory, Wright-Patterson AFB, Ohio, April 1973.

Lane, N. E. The influence of selected factors on shrinkage and overfit in multiple correlation. Naval Aerospace Medical Research Laboratory, Naval Aerospace Medical Institute, Pensacola, FL, September 1971.

Lees, M. A., Kimball, K. A., Hoffman, M. A., and Stone, L. W. Aviator performance during day and night terrain flight. USAARL Report No. 77-3, U. S. Army Aeromedical Research Laboratory, Fort Rucker, AL, December 1976.

Locke, E. A., Zavala, A., and Fleishman, E. A. Studies of helicopter pilot performance: II. The analysis of task dimensions. Human Factors, Vol. 7. No. 3, June 1965.

Northrop Corporation. Experiments to evaluate advanced flight simulation in air combat pilot training (6 volumes). Contract N62269-74-C-0314 (NADC). Northrop, Hawthorne, CA, March 1976.

Obermayer, R. W. and Vreuls, D. Measurement for flight training research. Proceedings of the 16th Annual Meeting of the Human Factors Society, Beverly Hills, CA, October 1972.

Obermayer, R. W., Vreuls, D., Muckler, F. A., Conway, E. J., and Fitzgerald, J. A. Combat-ready crew performance measurement system study. Contract F41609-71-C-0008 (AFHRL/FT). Manned Systems Sciences, Inc., Northridge, CA, May 1972.

Roscoe, S. N. Aviation Psychology. Ames, Iowa State University Press. In press.

Sanders, M. G., Kimball, K. A., Frezell, T. L., and Hoffman, M. A. Aviator performance measurement during low altitude rotary wing flight with the AN/PVS-5 night vision goggles. USAARL Report No. 76-10, U. S. Army Aeromedical Research Laboratory, Fort Rucker, AL, December 1975.

Schori, T. R. and Tindall, J. F. Multiple discriminant analysis: A repeated measures design. Virginia Journal of Science, Vol. 23, 1972, pp 62-63.

Smode, A. F. and Meyer, D. E. Research data and information relevant to pilot training. Volume I. General features of Air Force pilot training and some research issues. AMRL-TR-66-99, Vol. I, July 1966.

Vreuls, D., Obermayer, R. W., Goldstein, I., and Lauber, J. W. Measurement of trainee performance in a captive rotary-wing device. NAVTRAEQUIPCEN 71-C-0194-1. U. S. Naval Training Equipment Center, Orlando, FL, July 1973.

Vreuls, D., Wooldridge, A. L., Obermayer, R. W., Johnson, R. M., Goldstein, I., and Norman, D. A. Development and evaluation of trainee performance measures in an automated instrument flight maneuvers trainer. NAVTRAEQUIPCEN 74-C-0063-1. U. S. Naval Training Equipment Center, Orlando, FL, May 1976.

Waag, W. L. and Knoop, P. A. Planning for aircrew performance measurement R&D: U. S. Air Force. Proceedings of the Symposium on Productivity Enhancement: Personnel Assessment in Navy Systems, Naval Personnel Research and Development Center, San Diego, CA, October 1977.

Waag, W. L., Eddowes, E. E., Fuller, J. H., and Fuller, R. R. ASUPT automated objective performance measurement system. AFHRL-TR-75-3, Air Force Human Resources Laboratory, Williams AFB, Arizona, 1975.

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, Vol. 83, No. 2, 1976, pp 213-217.

Wooldridge, A. L., Breaux, R., and Weinman, D. Statistical issues in the use of multivariate methods for selection of flight simulator performance measures. NAVTRAEQUIPCEN 75-C-0091-1, U. S. Naval Training Equipment Center, Orlando, FL, in press.

ABOUT THE AUTHORS

Donald Vreuls has a B.S. in Psychology from the University of Illinois, an M.S. in Experimental Psychology from Trinity University and doctoral studies in Physiological Psychology at the University of Texas. He has 16 years of experience performing inflight and simulator research on aircrew performance and performance measurement methods on 18 Government Prime contracts with the U. S. Navy, U. S. Air Force, U. S. Army and the Department of Transportation. He has served as the Principal Investigator on 13 of these contracts, and has been employed by the Martin Company, the Bunker-Ramo Corporation, Manned Systems Sciences, Inc., and Canyon Research Group, Inc., where he is an Executive Scientist and Vice-President. He has authored or coauthored 41 technical reports, journal articles and professional presentations. He is an active private pilot with over 1400 hours of pilot in command and 380 hours of instrument time.

Lee Wooldridge has a B.A. in Psychology, a B.S. in Industrial Engineering and is completing his M.S. in Industrial Engineering at Florida Technological University. He has five years of experience in military training research, computer programming and measure selection statistical method development and application. He has originated many of the refinements to present multivariate measure selection techniques, has developed higher-order adaptive training system controllers, and has served as Principal Investigator on flight simulator studies of adaptive training logics. He has served on seven U. S. Government Prime contracts while at Florida Technological University as a research assistant, the Martin-Marietta Corporation as a Quality Engineer and Canyon Research Group, Inc. as a Staff Engineer. He has authored or coauthored six technical reports, and is a member of the AIAA.

MEASUREMENT OF THE SHIPBOARD PERFORMANCE CAPABILITIES OF NAVY ENLISTED PERSONNEL

Edward J. Pickering
Adolph V. Anderson

Navy Personnel Research and Development Center
San Diego, California 92152

ABSTRACT

This paper briefly describes some of the procedures the Navy uses to determine the degree to which Navy enlisted personnel are capable of performing the critical aspects of their jobs; these include the Personnel Qualification Standards (PQS) program, feedback programs of Navy training commands, the Personnel and Training Evaluation Program (PTEP), Training Readiness Evaluations, and Propulsion Examining Boards. Then, various experimental efforts are described in which job performance tests were administered to groups of Fleet personnel. Finally, a proposed "performance proficiency assessment system" is briefly described.

INTRODUCTION

In 1967 Earl Alluisi wrote:

> "Performance assessment" is one of the most important and difficult areas of current research. It is important in its own right, as any supervisor who has been called upon to justify the ratings of his workers can attest. It is important because it is the crux of the "criterion problem" for so much other work: the final validation of selection and training techniques depends upon the assessment of the performance of men who have been differently selected and trained. The final validation of an improved, human engineered, man-machine system depends upon it . . . . The assessment of man's behavior in the meaningful performance of complex tasks has challenged physiologists, engineers, and psychologists for many years. The task has been recognized as a difficult one; the problem has been formidable; and the solutions have been ephermeral . . . . Considerable quantities of good and respectable research have been published . . . . This research has advanced science generally, but it has failed to provide any significant progress toward performance assessment . . . . (pp. 375-376)

We think this would be a good and defensible statement for 1977, but let's take a closer look at some of the performance evaluation activities that have gone on in the Navy in the past and that are currently being pursued.

PRECEDING PAGE BLANK-NOT FILMED

In Dewey Stuit's book (1947), which summarizes the Bureau of Naval Personnel's personnel research during World War II, only one major attempt to relate selection and training information to shipboard performance is reported. That study, described by Bechtoldt, Maucker, and Stuit (1947) as "the first systematic study undertaken by the Bureau of Naval Personnel of the relationship between classification and training data and shipboard performance . . ." (p. 405), attempted to evaluate the performances of 1,868 men in six ratings on 27 ships shortly before the end of World War II. Bechtoldt, et al. considered such available information as quarterly marks and speed of advancement in rating as possible measures of proficiency, but rejected them because they decided that these measures were not sufficiently discriminative and were dependent upon many irrelevant administrative factors. They were looking for an overall measure of the quality of shipboard performance and, more specifically, a measure of technical competence. They decided not to administer objective achievement measures because of administrative difficulties, differences between ships, and a lack of time. After some preliminary study, they decided upon order of merit rankings in three areas: petty officer qualities, technical competency, and overall desirability. Two petty officers ranked each man in the sample. They found that the intercorrelations among the three measures were in the .80s and .90s and tended to be as high as the reliability of the separate rankings would permit. Therefore, they decided to use the technical competence rating alone for further analyses. Separate rank order ratings were obtained for each rating group on each ship. In one analysis, performance ratings were correlated with basic test battery scores and then averaged for the groups within each of the six Navy ratings. Of the 36 coefficients obtained in this analysis, 1 was in the .40s, 24 were in the .20s and .30s, and 11 were less than .20. One of the analyses of the relationships between school attendance and shipboard performance showed a substantial relationship between rank in class for basic technical training and the shipboard criterion measure. However, in their summary remarks, the authors concluded the following:

> Since a high correlation was found to exist between the technical competence, petty-officer qualities, and overall desirability ratings, there is grave doubt about the acceptability of the criterion. Until other criterion data are available, such as would be obtained from valid performance tests, it would be unwise to formulate any definite conclusions about the effectiveness or non-effectiveness of school training. (p. 408)

Now let's get a little more up to date and look at some of the procedures the Navy uses to "measure" the degree to which Navy enlisted personnel can perform the critical aspects of their shipboard jobs.

SOME EXAMPLES OF HOW THE NAVY "MEASURES" PERFORMANCE

Personnel Qualification Standards (PQS)

The Personnel Qualification Standards (PQS) program was developed to provide a procedure that could be used aboard ships to assure that officer and enlisted personnel are "qualified" to perform their assigned duties. These standards consist of written compilations of knowledges and skills required to qualify for a specific watchstation, to maintain a piece of equipment or a system, or to perform as a team member within an assigned unit, such as a highline detail or a damage control party. CNET Instruction 3500.3 states that:

> A watchstation, as it applies to PQS, refers to those positions normally assigned by a watchbill, usually of 4-hour duration, and, in the majority of cases, operator oriented. Maintenance refers to those tasks which pertain to technical upkeep of equipments and systems. Performance as a team member within the unit refers to those collections of knowledges and skills appropriate for standardized qualification which are not peculiar to a specific watchstation or equipment, but apply more broadly within the unit. A PQS is in the format of a qualification guide, which asks the questions a trainee must answer to verify his readiness to perform a given task and provides a record of his progress and final certification.

As directed by the Chief of Naval Education and Training and in response to requirements established by the Chief of Naval Operations, these "standards" are developed by the Personnel Qualifications Standards Development Group, Service School Command, San Diego.

At present, there are over 300 different PQS qualification guides available to the Fleet; a number that should eventually increase to over 1000. A typical qualification guide contains well over 1000 items on which various members of the crew must be qualified, either by attending courses, answering oral questions, or actually performing a specific task. This extremely large number of items is included because each guide attempts to list almost all of the things individuals have to know or be able to do to properly perform a specific set of duties. No attempt is made to distinguish between critical and noncritical knowledges and skills.

Illustrative items from the Personnel Qualification Standard for Damage Control are:

1. Describe the compartment numbering system used onboard your ship.

2. Define radiation dose rate.

3. Demonstrate your ability to bandage fractured ribs.

4. Show or describe the physical locations of the major Ship's Service Telephone stations on the ship and indicate the reason for each location.

5. Discuss the safety precautions that must be observed and perform the steps required to inspect piping and valves in your assigned space.

From the standpoint of providing reliable, valid information concerning the performance capabilities of shipboard personnel, PQS appears to have two major deficiencies. First, because of the vast number of PQS requirements that a given ship is responsible for, plus the many other duties and responsibilities of shipboard personnel, it is very unlikely that a sufficient amount of time or effort will be expended to assure that individuals are thoroughly proficient on those PQS items that they must be qualified on. As a result, "gundecking" is likely to occur; consequently, the validity of the information provided is questionable.

Second, the lack of specific evaluation guidelines makes it extremely unlikely that, for any given knowledge or skill, different raters will utilize the same criteria when judging performance. The system does not appear to provide performance information related to individuals and teams that is either objective or reliable.

Feedback Programs of Navy Training Commands

Both the Chief of Naval Education and Training (CNET) and the Chief of Naval Technical Training are very much concerned with developing improved procedures for gathering information from the Fleet concerning the adequacy of the training provided by various Navy training courses. In the future, these procedures may be changed considerably. Those currently in use have been described by Hall, Lam, and Bellomy (1976), who point out that although CNET Instruction 1540.3, Appraisal and Improvement of Training, requires both internal and external appraisals of training courses, it does not address external evaluation in detail. Rather, it describes the external evaluation process as being conducted to determine both how well course graduates can perform the job and the degree to which course learning objectives are relevant to requirements of this job. This information would be obtained through the use of graduate questionnaires, ship visits, task analysis, and letters from Fleet Commanding Officers.

A proposed revision to this instruction would formalize procedures for obtaining feedback from the Fleet concerning training effectiveness. It advocates the use of mailed questionnaires as the principal method of obtaining training feedback information. The recommended basic questionnaire form, which was developed by CNET's Training Analysis and Evaluation Group (Dyer, Ryan, & Mew, 1975a, b), consists of a listing of tasks taught in a specific course. Respondents are asked to evaluate each item in terms of (1) frequency of performance and (2) adequacy of school training for the task. Points on the frequency scale are: (1) never performed, (2) seldom performed or only in emergencies, (3) performed monthly, (4) performed weekly, and (5) performed daily. Points on the adequacy scale are: (1) task requires much more emphasis in school, (2) training less than adequate for task, increase emphasis, (3) training adequate for task, and (4) training more than adequate for task, reduce or eliminate training for this task. Responses would be obtained from samples of (1) graduates of a specific course who had been in the Fleet 6 months or less and (2) the supervisors of such graduates (Dyer et al, 1975).

Hall, Rankin, and Aagard (1976) voiced their concern about this reliance on questionnaires when they stated that:

> . . . the evaluation of training has not been given the attention and resources which are required for maintenance of a high-quality training system. . . . Present and planned procedures for obtaining and using training effectiveness information are not optimum. They may fail to yield the information needed for informed decision making about training. . . . The trend toward the exclusive use of questionnaires for obtaining data about training effectiveness reflects a lack of familiarity with other techniques that may be better suited for obtaining effectiveness information. A substantial number of methodological options are available for obtaining

such information. This selection and use should be based on consideration for specific elements of the evaluation situation. (pp. 39 and 40)

The value of using questionnaires for obtaining valid, reliable judgments concerning the need for increasing or decreasing the emphasis that should be placed upon tasks in a course curriculum can certainly be questioned. Additionally, even if such questionnaires prove to be adequate for that purpose, they will not, in the opinion of the authors of this paper, provide adequate information concerning the actual performance capabilities of Fleet personnel.

It should be pointed out that CNET has recognized the need for developing other methods for gathering training feedback information; its Training Appraisal and Surface Warfare Division (N8) is currently investigating other methods for improving the training feedback process. However, current budgetary and personnel constraints preclude the immediate development of an all-encompassing, centrally controlled training appraisal system. Also, because of the need to minimize programs that infringe on Fleet time and assets, their investigation is concentrating on identifying existing programs that can provide the desired feedback with little or no modification.

The Personnel and Training Evaluation Program

At the present time, one of the more significant evaluation programs within the Navy is the Fleet Ballistic Missile Weapon System's Personnel and Training Evaluation Program (PTEP). PTEP is responsible for assessing the skill levels of Fleet Ballistic Weapon System personnel when they are undergoing training and when they are performing duties aboard submarines. The majority of submarine personnel in applicable ratings are tested twice a year (Hall, Lam & Bellomy, 1976; Braun & Tindall, 1974). The PTEP achievement testing program employs two types of test instruments:

1. Four-alternative multiple-choice tests are used to obtain information concerning knowledges required to carry out job duties.

2. Paper-and-pencil skill exercises are used to "measure skill performance in terms of skill related knowledge application" (Braun & Tindall, 1974, p. 1).

The paper-and-pencil skill exercises consist of equipment operation or maintenance problems that the examinee must solve by identifying the procedural steps that should be taken to perform the particular equipment operation or to remedy the apparent equipment malfunction. These tests are related to the early "Trainer-Tester" devices developed by Van Valkenburg, Nooger, and Neville, Inc. However, for a given set of conditions, the "Trainer Tester" allows free exploration of test points, pins, contacts, etc. Therefore, multiple-path solutions exist when using this device. In the PTEP skill exercises, only four selections are allowed for each major step in a single-path problem-solving sequence. The PTEP exercises utilize latent image (invisible ink) printing for solution masking and response feedback (Laabs, Panell, & Pickering, 1977).

Braun and Tindall, in discussing these exercises, state:

Although each exercise may not necessarily reflect the exact procedure which the examinee would prefer to use,

> it is expected that a trained technician would recognize the exercise solution path as being valid and the most appropriate one from among those choices made available to him. His ability to recognize and logically work through the exercise solution path is considered to be directly indicative of his ability to apply the various diagnostic skills required of him when encountering the several different possible situations of equipment operation and maintenance . . . . The particular procedural steps which make up the exercise solution are selected as necessary to fulfill the defined testing objectives of the exercise with respect to testing specific skills normally required of the examinee . . . . The exercise solution is supported by realistic pictorials of control panel indications, duplications of actual equipment printouts, photographs of waveforms as observed with an oscilloscope, and the same documented procedures, technical data, and equipment maintenance diagrams as would normally be available in the real environment of equipment operation and maintenance. (pp. 4 and 5).

It should be emphasized that the PTEP skill exercises have not been empirically validated against real measures of performance on the actual equipment. In the absence of such information, results from this testing program should be used with care. This appears to be especially true if one considers what previous research concerning the validity of tests of this type has shown. Crowder, Morrison, and Demaree (1954) reported correlations of .12 and .16 between two forms of a trainer-tester test and a performance test on the actual equipment. Steinemann (1966) found correlations that ranged from -.50 to +.14 between various measures yielded by a trainer-tester test designed to measure ability to troubleshoot a superhetrodyne receiver and comparable measures from a test on the actual equipment. Steinemann points out that, in the simulated troubleshooting, test measures are obtained by simply erasing the covering material. Consequently, the measure given is always accurate. In actual practice, checks and measurements require considerably more effort and the accuracy of the reading depends upon the skill with which the test equipment is used. Steinemann found that, when taking the actual performance test, subjects often repeated the same check or measure because they were uncertain of the accuracy of their findings. He states:

> Dubious measurements tended to affect the entire troubleshooting sequence. Reliance upon an incorrect reading, for example, could lead examinees to a false casualty assumption. Conversely, uncertainty over a correct reading sometimes caused students to persist in repeating an unproductive line of troubleshooting strategy . . . . In the actual task, students were reluctant to unsolder or disconnect components from the chassis, but in the simulated task, where parts replacement required virtually no effort, students too often resorted to parts replacement in an effort to solve the problem. (pp. 10-11)

Steinemann concluded that the evidence strongly suggests that caution should be exercised in assuming that any simulated performance measure, even when it has considerable common identity to the actual task, will provide a valid estimate of proficiency on the actual equipment.

Shriver and Foley (1974) compared performance on a series of maintenance tests with performance on a parallel set of specially developed paper-and-pencil "graphic symbolic substitute" tests. When these symbolic substitute tests were being developed, it was hypothesized that they would be more valid than previously developed symbolic tests because they would contain more realistic task "clutter." For example, in testing troubleshooting, if the subject wanted to know the voltage at a specific test point, he would be provided with a picture of a voltmeter that displayed the requested information in the same way he would see it on the actual equipment, rather than being provided with a printed voltage readout. The symbolic tests were subjected to a small-scale validation in which novice technicians took both the symbolic and performance versions of the tests. On the basis of these results, the symbolic troubleshooting test was modified and then subjected to validation utilizing experienced technicians. The results of both validations are summarized in Table 1. It was concluded that the symbolic tests, with the exception of the one on soldering, showed sufficient promise to justify further consideration and refinement. However, the authors point out that valid symbolic substitute tests cannot be developed for any job activity until good job-performance tests are available (Shriver & Foley, 1974). Unfortunately, at the present time, the Navy has no mechanism for the development and administration of job-performance tests. Foley (1975) points out that, in his opinion, symbolic substitutes of high empirical validity can be produced; but that such tests will never eliminate the requirement for the liberal administration of actual job performance tests to maintenance personnel: "We can never include all aspects of an actual performance of a task in a paper-and-pencil symbolic representation of that task . . . ." (p. 7). The authors of this paper wholeheartedly agree. Furthermore, paper-and-pencil tests undoubtedly have the same limitations when they are utilized to measure operator skills.

Table 1

Summary of Relationship Found Between Performance and Symbolic Versions of Electronic Maintenance Tests

Test Area	Number	Phi Coefficient	Tetrachoric Correlation
Novice Subjects			
Checkout	4	1.00	-
Removal and Replacement	14	.43	-
Soldering	4	0	-
General Test Equipment	6	.67	-
Special Test Equipment	6	.33	-
Alignment and Adjustment	19	.58	-
Troubleshooting	9	-.33	-
Experienced Subjects			
Overall Troubleshooting	30	.47	.68
Chassis Isolation	30	.73	.81
Stage Isolation	30	.33	.46
Piece/Part Isolation	15	.07	.16

Note. From Shriver and Foley, 1974.

Performance of LAVA Operators

One example of a performance test that is routinely administered in the Fleet is a test of the physical analysis skills of LAVA operators. LAVA is a shipboard sonar receiving set that is an integral part of the LAMPS/LAVA ASW Surveillance System. In this system, passive sonar signals are received from sonobouys by a LAMPS equipped heliocopter; the heliocopter then transmits these signals to a LAVA equipped destroyer. The incoming sonar signals are printed out on grams that are analyzed for contact signature characteristics by specially trained operators.

The Fleet Aviation Specialized Training Group, Pacific (FASOTRAGRUPAC) utilizes recordings of actual sonar contacts in conducting weekly training exercises. These recordings are broadcast from a "Rooftop Trainer" on North Island, San Diego, to all LAVA equipped ships in port. A LAMPS configured heliocopter serves as the data link between the transmitter and the LAVA receiving sets. Each exercise consist of six tape cuts. The first cut is used for equipment checkout of the LAMPS/LAVA euqipment and the remaining five cuts are of actual contacts. Following each cut, ships report their classification of the contact. After the exercise, each ship is given its percentage of correct classifications and this information is also transmitted to the Commander Surface Forces, Pacific. However, no diagnostic information is provided concerning the types of errors that were committed.

Training and Readiness Evaluations

The Commander Naval Surface Force, U. S. Pacific Fleet, has outlined the requirements for training exercises and inspections that are designed to establish and maintain maximum battle readiness. No attempt will be made here to describe these requirements in any detail; however, a few words will be devoted to discussing the general purpose of these evaluations.

The evaluation of battle readiness is administered to accomplish readiness objectives in as flexible a manner as possible and to minimize formal reporting. Operational Readiness Evaluations (OREs) are designed to analyze the overall readiness of ships in meeting mission and operational requirements. These evaluations consist of two separate but closely related components: battle problems and selected readiness exercises. The primary purpose of the battle problem is to provide a medium for evaluating the ability of all departments on a given ship to function together as a team in simulated combat operations and to accomplish the tasks required by the problem. All tasks cannot be incorporated into a single battle problem; consequently, those that can best be evaluated by separate exercises and that do not require simultaneous actions by several departments are not considered to be part of the battle problem. Those required operational capabilities that are not evaluated during the battle problem, or for which additional evaluation is desired, are examined through the completion of selected readiness exercises. All OREs involve evaluations and grades (i.e., Outstanding, Excellent, Good, Satisfactory, or Unsatisfactory); however, in order to minimize operational requirements, detailed evaluation reports of individual exercises do not have to be forwarded to the Type Commander.

OREs are designed to measure battle readiness of a ship; consequently, in general, detailed information concerning specific performance deficiencies of team members is not gathered.

1200 PSI Propulsion Examining Boards

In the past, the material readiness of 1200 psi propulsion plants was not completely satisfactory, as evidenced by the fact that these plants were involved in an inordinate number of serious personnel and/or material casualities. Personnel error, attributable to insufficient training in plant operation and maintenance, was identified as a major cause of these casualties. As a result, it was decided that a formal procedure was needed for examining 1200 PSI ships to ascertain the state of training of propulsion plant personnel and to determine the material condition of their propulsion plants. Consequently, OPNAV Instruction 3540.4 established 1200 PSI Examining Boards on the staffs of the Commanders in Chief, U.S. Pacific Fleet and U.S. Atlantic Fleet. In 1976, in OPNAVINST 3540.4B, the authoritiy of these boards was extended to include the examination of some 600 PSI steam powered ships and ships equipped with LM-2500 gas turbine propulsion plants:

> These boards provide a formal means of examining conventionally powered ships to ascertain the state of training and qualification of propulsion plant personnel, and to determine the material condition of the propulsion plants. In addition they evaluate the capability of the propulsion plant to meet existing readiness standards and the capability of ship's engineering personnel to operate the propulsion plant properly and safely.

Each board consists of a minimum of 17 officers who have had appropriate Fleet experience and training. The boards conduct two types of examinations: (1) Initial Light-Off Examinations and (2) Operational Propulsion Plant Examinations.

Initial Light-Off Examinations (LOE) are conducted before the first light-off of any boiler or any gas turbine during a regular overhaul, major conversion, or restricted availability in excess of 4 months. The board interviews shipboard personnel, administers written tests, reviews administrative procedures including training programs, and inspects the system. Propulsion plant drills are not formally required during a LOE. However, "simple evaluations" (such as the ability to sample and analyze boiler water and feedwater) can be conducted at the discretion of the senior member of the examination team.

Operational Propulsion Plant Examinations (OPPEs) are conducted no more than 4 months after the end of a regular overhaul, restricted availability in excess of 6 months, or post shakedown availability for new construction ships or ships having undergone major conversions. Subsequent examinations are administered at approximately 18-month intervals. During an OPPE, the examination team checks safety devices, administers written and oral examinations, observes personnel as they operate equipment at sea under a variety of situations, and observes personnel as they take part in a fire drill.

These examinations result in a judgment by the board concerning the personnel and material readiness of the ship in question. Specific deficiencies are identified and these must be corrected before a ship is considered to be ready for unrestricted operation in the case of an OPPE or light-off in the case of a LOE. Ships are reexamined by the Propulsion Examining Board to assure that appropriate corrective actions have been taken.

Shipboard Performance Testing

Harris and Mackie (1962) described the status of shipboard performance testing of individuals as it existed in the first part of 1962. In 1977, that status does not appear to have changed substantially. Harris and Mackie interviewed 233 supervisors aboard ships and at other operational commands. They found that, excluding written examinations for advancement in rating, performance evaluations were based primarily upon subjective impressions gained from a man's performance on the job. However, the limited degree to which job performance tests were being used was not consistent with the generally positive attitudes of supervisors toward such tests. Although they were being used by only 17 percent of the supervisors surveyed, 72 percent of those supervisors thought that performance tests reflected performance capabilities very well. On the other hand, 97 percent were using supervisor's judgments to assess performance, but only 63 percent thought such judgments did a good job of reflecting performance capabilities. Almost two-thirds of the individual performance testing that was being done aboard ships involved the radiomen rating; this was the only rating among those studied where performance tests were available that had been developed by the Navy Examining Center for use in conjunction with written advancement in rating examinations.

The two primary reasons given for not using performance tests were that (1) such tests were not suitable or practical for the rating in question and (2) suitable tests had not been developed or were not available.

Before concluding this section, it should be emphasized that the measurement efforts that have just been discussed do not represent an exhaustive list of measurement efforts that take place on Navy ships. Also, such programs come and go and they change rapidly. However, the programs described are believed to be fairly representative of the major types of evaluations that are currently a part of the sailor's life.

EXPERIMENTAL PERFORMANCE MEASUREMENT EFFORTS

In the past, NPRDC and others have developed diagnostic, job performance tests that were then administered to groups of Fleet personnel in order to investigate various training problems. Results of their administration to samples of Fleet personnel have consistently demonstrated the value of such tests as a tool for detecting skill deficiencies. In the paragraphs that follow, some examples of these testing programs will be described briefly. First, a few words will be said about a study that investigated the general problem of evaluating shipboard performance.

Methods of Measuring Shipboard Performance

In 1954, Wilson, Mackie, and Buckner carried out an extensive investigation of methods for evaluating the shipboard performance of Navy enlisted personnel. This investigation involved the development of four types of measurement instruments (i.e., performance rating scales, performance checklists, written job knowledge tests, and practical performance tests). Since the study has been described in detail in a series of reports (Wilson & Mackie, 1952; Wilson, Mackie, & Buckner, 1954a, 1954b, 1954c), further discussion is not needed here; however, a number of findings should be mentioned. In fact, if the study were to be repeated in 1977, 23 years later, the findings would probably be about what they were in the original study.

The researchers found that the correlations between actual job performance tests, written tests, and rating scales were low to moderate. They indicated that their results suggested that officers and petty officers did not have valid information about the specific capabilities of the men that work for them. These officers and petty officers may have had reasonably valid general impressions concerning their men's abilities; however, they could not have been expected to know whether or not a given individual can perform a specific task. Consequently, the results suggested that rating devices would be more useful if they were oriented toward the general characteristics of shipboard performance rather than toward specific tasks men might be required to perform on the job. The low correlations between written and performance tests indicated that (1) verbal understanding of a task does not necessarily imply the ability to perform that task, and (2) ability to perform a task does not necessarily mean that an individual can answer written questions about it. The authors concluded that shipboard performance measures in the Navy should be designed to obtain indications of both knowledge related to a specific job and the ability to perform the practical work required by that job (Wilson, et al., 1954c). As regards the problems of implementing performance testing procedures, they observed that:

> Since potential performance test users were found to agree generally with performance criteria specialists that performance tests are a valuable, and often the best, method of determining a man's capability for performing the critical tasks of his job, the primary problem in implementing the use of performance tests becomes not one of selling the idea of their value, but one of overcoming the practical barriers limiting their feasibility.
>
> The problem of feasibility can be resolved by either changing the environment in which tests are to be used, by developing tests acceptable and useable in the existing environment, or both. Changing the environment would call for an increased emphasis by Navy commands on demonstrated performance capability through use of performance tests. In contrast to the limited necessity for selling supervisors and instructors on the value of performance testing, such a change would require considerable indoctrination of command personnel into the benefits of performance testing, involving, as it would, the expenditure of personnel and equipment time at the apparent sacrifice of some short-term operational or training goals.
>
> Resolving the problem by providing practical performance tests "packaged" to be acceptable and useable in the existing environment, while at the same time maintaining adequate standards of test reliability and validity, is a formidable challenge for the test developer. If it is difficult to meet this challenge within the constraints imposed by the training environment, it will be considerably more difficult to meet it within those imposed by the operating environment. (pp. 11 and 12)

Maintenance of the AN/URC-32 Transceiver

In a study conducted by Rigney and Fromer (1965), they observed that, over

the years, very little criterion data had been collected by researchers that involved actual performance testing. They did not find this surprising in view of the many difficulties related to the collection of such data. The major difficulties being:

> 1. The collection of data on the actual performance of electronic maintenance tasks is very time consuming. It may take a full day or longer to obtain a good sample of data on a single technician. A relatively large number of technicians, (e.g., 50 to 100) must be tested on a single piece of equipment or system. Testing should be done on a number of electronics systems, at least one of each major category, if results are to be generalized to the entire electronic equipment complement of a ship.
>
> 2. A relatively controlled testing situation is difficult to achieve. It is desirable to have full use and control of the electronic system being used as the test vehicle, since the system would have to be used on a daily basis solely for testing purposes. In addition, good logistic and maintenance support is required to keep the system in peak operating condition, and it is essential to have a full set of test equipment associated with the specific system.
>
> 3. It is often difficult to obtain the appropriate subjects . . . for testing purposes. This is due to a reluctance of maintenance officers to release their maintenance personnel for an entire day or longer, transportation problems, and sometimes an apprehension, on the part of the people being tested, of being evaluated on a personal basis. The management of subject scheduling is an additional problem.
>
> 4. It is difficult to obtain an expert observer. The observer must be expert both in data collection and as an electronics technician, especially on the system being used as a test vehicle. (pp. 3 and 4)

Rigney and Fromer (1965) were able to overcome most of these difficulties when they collected criterion information on electronic technician skills in performing corrective maintenance of the AN/URC-32 transceiver. The fact that it takes a considerable amount of time to administer performance tests was accepted as part of the cost of the project. However, the testing time was held down to 1 day in order to avoid the problem of recalling subjects. The problem of getting a reasonable sample of technicians to a suitable, well controlled testing site was solved by setting up the testing vehicle in a building within the Long Beach Navy Base where it was within walking distance of destroyers, cruisers, and tenders. The expert observer problem was solved by obtaining a maintenance expert and training him in the fundamentals of observation.

This study utilized a system for classifying corrective maintenance tasks that had been developed previously by Rigney, Cremer, and Towne (1965). Four tests were developed that measured performance on four major parts of this classification scheme: system state recognition, localization, circuit isolation, and component isolation. The system state recognition test measured the

ability of technicians to recognize whether or not the transceiver was functioning correctly in each of its ten modes, and the localization test measured the technician's ability to utilize symptom information and localize a problem to a "fault area" within the system. The circuit and component isolation tests measured the ability of technicians to isolate the trouble to one of several possible faulty circuits or to a single component within a faulty circuit, respectively.

The test was administered to 54 Fleet electronic technicians, all of whom were responsible for the maintenance of the AN/URC-32 on their ship. The test results showed that very few of the technicians tested could successfully perform the system state recognition and fault localization tasks. They were somewhat better at circuit and component isolation; however, their success was highly dependent upon the nature of the malfunction. According to Rigney and Fromer, the major weaknesses in specific skills were:

> (1) an inability to operate the equipment properly in all of the modes of operation, accompanied by a poor understanding of the less frequently used modes, (2) an inability to abstract information from technical manuals regarding the relationships between circuitry and front panel indications, or a lack of prior knowledge concerning these relationships, and (3) an inability to use test equipment properly and efficiently. (p. 62)

The researchers were able to make a number of training, hardware design, and software recommendations for improving corrective maintenance of the AN/URC-32.

Proficiency of JEZEBEL Operators

At the request of the Chief of Naval Operations (OP-56), NPRDC carried out an extensive testing program designed to obtain information concerning the gram-analysis proficiency of JEZEBEL operators. With the aid of a group of JEZEBEL instructors, alternate forms of a job-sample test were developed. These tests were administered in VS and VP squadrons early in 1969, in mid-1969, and early in 1970. A total of 1,749 useable returns were obtained. The test results not only identified a number of specific performance deficiences, but also showed that proficiency decreased markedly over time because the operational training structure had no provisions for systematic refresher or review training. As a result of these findings, self-study training materials were introduced into the squadrons, and a systematic gram analysis procedure was developed that was designed to eliminate many of the types of common errors that the testing program identified. An evaluation of these materials indicated that their Fleet-wide use could increase classification accuracy by as much as 20 percent.

Submarine Sonar Operator Classification Skills

In 1968, Mackie, Parker, and Dods of Human Factors, Inc., carried out what was probably the first comprehensive attempt to measure the classification ability of submarine sonar personnel. They discovered that there were very extensive individual differences among submarine sonar personnel in their ability to classify sonar contacts. To a considerable extent, these differences were associated with amount of operational experience. The results suggested that considerable improvement in the average level of Fleet performance was possible through more extensive initial training or more frequent refresher training.

As a result of this study, Mackie (1968) recommended that the classification skills of all submarine sonarmen, supervisors, and instructors be measured at least once a year by means of objective performance tests. He further suggested that advancement in rate be made, in part, contingent upon satisfactory performance on classification tests. In 1972, the National Security Industrial Association Report on ASW (Buaas, 1972) observed that the requirement for the development of standardized performance tests for assessing progress and operational readiness of sonar operators, sonar officers, and ASW teams had still not been met. Although this had again been determined to be an urgent requirement. Even closer to the present, in 1974 Mackie observed that to his knowledge neither of his earlier recommendations had been implemented although the need seemed greater than ever (Mackie, 1974). Today, in 1977, this requirement has still not been met.

Building on Mackie's work, NPRDC carried out research concerning the ability of submarine sonar operators to classify sonar contacts by means of sonar information presented both in the form of sound frequency grams and aurally. Using a testing format that was based upon an analysis of the classification task aboard a modern nuclear submarine, data was gathered on the ability of operators (1) to recognize and identify relevant gram signature characteristics and audio cues, (2) to integrate this information, and, (3) finally to classify correctly various contacts. The test results pointed out specific performance deficiencies that were considered to be correctable by modifying the training procedures utilized at that time. The most important deficiency found was that personnel were not utilizing any specific system when integrating either aural or gram information. As previously mentioned, similar deficiencies were also found by Mackie. In an attempt to partially remedy this situation, the classification tasks of submarine sonar operators were analyzed and decision-action-diagrams (decision trees) were constructed to aid in the cue integration process. Evaluations of these procedures showed that a small amount of training, followed by use of the decision trees as job aids, resulted in significant improvements in the performance levels of Fleet personnel. For example, on one very critical set of tasks, the average pretest to posttest improvement was 36 percent.

Use of Electronic Test Equipment

As part of a survey of the duties, training, and proficiency of sonarmen, diagnostic performance tests were developed (Anderson & Pickering, 1959) that measured ability to use electronic test equipment (i.e., volt-ohm-meter, vacuum tube volt meter, oscilloscope, and signal generator). These tests were administered to 396 Pacific Fleet sonarmen. It was found that proficiency was much lower than anticipated; on the average, the ability of sonarmen at all grade levels was poorer than desirable. The diagnostic test format permitted the fairly precise pinpointing of the specific types of errors being committed. This diagnostic information resulted in the development of modified training procedures and marked improvement in the performance of "A" school graduates (Pickering & Abrams, 1962). It is interesting to note that, whenever technicians are tested on their ability to use electronic test equipment, in general, performance is found to be less than adequate. In the past, this has been true for surface sonar technicians (Anderson & Pickering, 1959; Branks, 1966), submarine Sonar Technicians (Anderson & Pickering, 1959; PTEP Evaluation, 1976), Electronics Technicians (Anderson, 1962), Fire Control Technicians (Bilinski, 1965), and Missile Technicians (Laabs, Panell, & Pickering, 1977). As a result of such testing, corrective measures are often taken; however, it does not appear that increases in proficiency are maintained over long periods of time. Perhaps what

is needed here is some procedure for periodically testing samples of Fleet electronic technicians on critical skills like the use of test equipment.

The Personnel Readiness Training Program

The Personnel Readiness Training Program explored the feasibility of utilizing diagnostic, performance-oriented tests to identify critical skill deficiencies and of developing individually prescribed shipboard training packages to correct them (Anderson, Laabs, Winchell, & Pickering, 1977; Laabs, Harris, & Pickering, 1977; Laabs, Panell, & Pickering, 1977; and Winchell, Panell, & Pickering, 1976). Prototype testing/training packages were developed for three areas of application: (1) maintenance tasks performed by Missile Technicians (MTs) on the Missile Test and Readiness Equipment (MTRE MK-7, MOD-2), (2) operator tasks performed by submarine Sonar Technicians (STs) on the AN/BQR-20A passive real-time frequency analyzer, and (3) operator and maintenance tasks performed by Boiler Technicians (BTs) on the 1200 PSI Steam Propulsion Plant.

The program's experimental design called for three groups of subjects in each area of application. These groups were a Control Group, a Diagnostic Feedback Group, and a Diagnostic Feedback + Training Group. All three groups were given a diagnostic pretest and, after an intervening period of approximately 5 months, they were given a posttest. For all groups, the time between pretest and posttest was occupied with regular duties. After the pretest, members of the Control Group were provided only with the overall percentage of test items they had performed correctly; they were not given any specific feedback. Members of the Diagnostic Feedback Group were provided with specific feedback information with respect to the performance weaknesses revealed by the pretest, but, like the Control Group, they were not provided with information on how to correct their deficiencies. Members of the Feedback + Training Group were given immediate feedback on their specific deficiencies and a specifically selected set of training materials designed to correct individual deficiencies. It was suggested that the time between pretest and posttest be partially occupied by working with the training materials.

The sampling unit for this study was a ship. Twelve ships were involved for each of the three areas of application: four ships for the Control Group, four for the Diagnostic Feedback Group, and four for the Diagnostic Feedback + Training Group. On each participating ship, all pertinent personnel were included.

For the BQR-20 application, two diagnostic tests were developed: a performance test and a written test. For the performance test, the BQR-20 was interfaced with an UNQ-7 tape recorder. Four test tapes provided inputs to the system--three contained simulated sonar signals and one contained active sonar transmissions. The test involved search, contact investigation, and tracking tasks; signal-to-noise ratio calculations; and a ping interception task. The written test contained 42 items covering knowledge considered essential to the effective operation of the BQR-20.

For the MTRE application, four diagnostic tests were developed: a preventive maintenance test, a corrective maintenance test, a simulated troubleshooting test, and a test equipment test. The preventive maintenance test consisted of three problems in which technicians were required to perform various maintenance checks on the actual MTRE equipment. The corrective maintenance test consisted of two troubleshooting problems on the actual MTRE equipment--one

involving a prefaulted module and the other, a misadjusted potentiometer on an amplifier card. This test was administered, in place of the preventive maintenance test, to the one MT on each submarine who had primary responsibility for MTRE during the submarine's last patrol. The simulated troubleshooting test was a paper-and-pencil test designed to measure the ability of the MT to logically apply his knowledge of equipment maintenance. For the test equipment test, a specially designed test signal generator was used to provide the electronic signals for the eight test problems. Problems involved use of the oscilloscope (USM-281), the differential voltmeter (John Fluke), and the volt-ohm-microammeter (Simpson).

Two diagnostic tests were developed for the 1200 PSI Steam Propulsion Plant application. The first of these was a multiple-choice test on basic mechanical skills and knowledges. This test described job situations and the questions involved many actions that simulated those occurring on the job; (e.g., identifying hand tools, reading measurement devices, recognizing equipment components, and interpreting charts, tables, and diagrams). The second test was a 28-item multiple-choice test on the use of the Engineering Operational Sequencing System (EOSS), which is a detailed system of operating procedures for the 1200 PSI Steam Propulsion Plant. If this system is understood and used correctly, it should result in safe steaming of the plant by propulsion engineering personnel. The test involved reading and interpreting a variety of EOSS documents that were reproduced in the test booklet. It should be noted that the only place it would have been feasible to carry out an actual performance test of propulsion plant procedures was aboard ships. Although the specifications for such a test were developed, it was not possible to get permission to do any performance testing aboard ships.

On the basis of job task analyses, remedial training materials were developed for each of the three applications. These materials were packaged in such a way that they could be assigned in relatively small modules to match areas of weakness revealed by the test.

A total of 63 MTs, 56 STs, and 163 BTs were tested. In general, it was found that providing testees with feedback alone did not result in improved performance, and that performance deficiencies can be corrected through remedial training if the training materials and procedures are used appropriately. From the standpoint of this paper, the fact that in all three areas significant performance deficiencies were found is of more interest. For example, in the AN/BQR-20A application, over all groups on the pretest, only 75 percent of the front panel and CRT controls were set correctly on the search task, 65 percent on the contact investigation task, and 67 percent on the tracking task. As described by Winchell, et al. (1976) this finding is of real concern to the Navy:

> While it is accurate to say that, under some conditions, all of the available contact information may be obtained even when certain front panel controls are set improperly, the danger is that, on other occasions, these improper settings may make it unlikely that vital contact information will be obtained. Further, in other more serious cases, certain settings on some switches make it impossible to obtain contact information. Additionally, overdriving the signal, a condition observed frequently during testing, may create spurious, system-generated artifacts that can mask real contact information or be misinterpreted as relating to a nonexistent contact. (p. 14)

In the MTRE application, the pretest results indicated that there were numerous serious deficiencies. In fact, the initial performance levels were so low that almost all MTs were assigned the entire remedial training package. Prior to testing, it had been anticipated that a considerable number of MTs would require little or no remedial training (Laabs, Panell, & Pickering, 1977). In the Steam Propulsion Plant application, the pretest results showed that there were many deficiencies in answering sets of performance-oriented written test items related to the operation and maintenance of the 1200 Pound Steam Propulsion Plant. The seriousness of the problem is indicated by the fact that the test only covered basic skills and knowledges that were considered to be essential for proper operation and maintenance of the plant (Laabs, Harris, & Pickering, 1977).

There were probably many reasons for the deficiencies that were found in the Personnel Readiness Training program; however, these reasons will not be discussed here. It is sufficient to note that serious deficiencies were found and that it is probable that comparable deficiencies exist in relation to many other job skills. The authors of this paper believe that the only way to detect such deficiencies is through some sort of programatic performance testing of samples of Fleet personnel.

Aviation Maintenance Skills

From the preceding paragraphs, it can be seen that, in general, when performance tests previously have been administered to groups of Fleet personnel, a number of deficiencies have been detected. However, it should be pointed out that this is not always the case. For example, Jones and Abrams (1960a, 1960b) administered performance tests to 61 Aviation Electronics Technicians (ATs) and 27 Aviation Structural Mechanics (AMs) who had just entered the Fleet upon graduation from Class "A" School. The same tests were also administered to 44 ATs and 27 AMs who had Fleet experience. The ATs were tested on use of test equipment, soldering procedures, troubleshooting procedures, and use of manuals and publications. The AMs were tested on interpretation of shop drawings, shop computations, and selection of materials and preparation for riveting, welding and cutting with a torch. In general, the test results showed that most "A" School graduates could perform in a satisfactory manner. The experienced men, however, performed at a considerably higher level.

A LOOK TO THE FUTURE

Where do we go from here? Since World War II, investigator after investigator has recognized the need for quality job performance data in order to properly evaluate the effectiveness of such personnel practices as selection and training. Unfortunately, however, it is very clear that useful job performance testing is difficult, expensive, and demands substantial expertise. It is evident that the cost of measuring all aspects of job performance for all Navy incumbents would be prohibitive. As one possible solution to these problems, NPRDC has initiated the development of what we call a performance proficiency assessment system. A fundamental notion underlying this development is that we will be attempting to evaluate the personnel processes such as selection, assignment, training, and on-the-job training rather than attempting to evaluate every individual who occupies a job of interest. Furthermore, we will concern ourselves only with critical and important skills rather than all skills. What we are proposing is a quality control approach similar to that used in the manufacturing of industrial products. This approach will require:

1. An identification of critical and important tasks.

2. The establishment of appropriate performance criteria relative to critical tasks.

3. The development and implementation of appropriate sampling procedures relative to both tasks and job incumbents.

4. The application of effective procedures for measuring the performance of job incumbents in quantifiable terms.

5. An understanding and quantification of the personnel processes that bring individuals to their jobs.

6. A capability for analyzing performance data and relating results to appropriate personnel processing practices.

7. A capability to provide personnel managers with appropriate and understandable reports.

For our development of a prototype system, we have selected two enlisted ratings for study: The Interior Communications Electrician and the Surface Sonar Technician. We start this effort with a full recognition that if we are to be even moderately successful we will have to overcome fairly horrendous obstacles.

REFERENCES

Alluisi, E. A. Methodology in the use of synthetic tasks to assess complex performance. Human Factors, 1967, 9, 375-384.

Anderson, A. V. Training, utilization, and proficiency of Navy electronics technicians: III. Proficiency in use of test equipment (Bureau of Naval Personnel Technical Bulletin 62-14). San Diego: Naval Personnel Rsearch Activity, September 1962.

Anderson, A. V., Laabs, G. J., Winchell, J. D., & Pickering, E. J. A personnel readiness training program: Final report (NPRDC Technical Report 77-39). San Diego: Navy Personnel Research and Development Center, August 1977.

Anderson, A. V., & Pickering, E. J. The proficiency of pacific fleet sonarmen in the use of test equipment (Bureau of Naval Personnel Technical Bulletin 59-30). San Diego: Naval Personnel Research Field Activity, November 1959.

Bechtoldt, H., Maucker, J., & Stuit, D. Prediction of performance of enlisted personnel aboard ship. In D. Stuit (Ed.), Personnel research and test development in the Bureau of Naval Personnel. Princeton, N. J.: Princeton University Press, 1947.

Bilinski, C. R. Training, utilization, and proficiency of Navy fire control technicians (Technical Bulletin 66-2). San Diego: Naval Personnel Research Activity, July 1965.

Branks, J. Proficiency of basic sonar maintenance trainees in the use of common test equipment (NPRA Research Report 66-14). San Diego: U. S. Naval Personnel Research Activity, January 1966.

Braun, F. B., & Tindall, J. E., Jr. A new sequential multiple-choice testing device. Norfolk, Va., Data Design Laboratories, October 1974.

Buaas, M. H. (Chair). Personnel, training, and human factors subcommittee ad hoc studies conducted for the seventh NSIA report on ASW. Washington, D.C.: National Security Industrial Association, February 1972.

Crowder, N., Morrison, E. J., & Demaree, R. G. Proficiency of Q-24 radar mechanics: VI. Analysis of intercorrelations of measures (AFPTRC-TR-54-127). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center, 1954.

Dyer, F. N., Ryan, L. E., & Mew, D. V. Procedures for questionnaire development and use in Navy training feedback (TAEG Report No. 201). Orlando, FL: Training Analysis and Evaluation Group, October, 1975. (a)

Dyer, F. N., Ryan, L. E., & Mew, D. V. A method for obtaining post formal training feedback: Development and validation (TAEG Report No. 19). Orlando, FL: Training Analysis and Evaluation Group, May 1975. (b)

Establishment of 1200 PSI Propulsion Examining Boards (OPNAVINST 3540.4). Washington, D. C.: Chief of Naval Operations, November 1972.

Foley, J. P., Jr. Criterion referenced measures of technical proficiency in maintenance activities (AFHRL-TR-75-61). Wright-Patterson Air Force Base, Ohio: Air Force Human Resources Laboratory, October 1975.

Hall, E. R., Lam, K., & Bellomy, S. G. Training effectiveness assessment: Volume I, Current military training evaluation programs (TAEG Report No. 39). Orlando, FL: Training Analysis and Evaluation Group, December 1976.

Hall, E. R., Rankin, W. C., & Aagard, J. A. Training effectiveness assessment: Volume II: Problems, concepts, and evaluation alternatives (TAEG Report No. 39). Orlando, FL: Training Analysis and Evaluation Group, December 1976.

Harris, D., & Macki , R. R. Factors influencing the use of practical performance tests in the Navy (Technical Report 703-1). Los Angeles: Human Factors Research, Incoprorated for Office of Naval Research, August 1962.

Interservice procedures for instructional systems development (NAVEDTRA 106A). Pensacola: Naval Education and Training Command, August 1975.

Jones, E. I., & Abrams, A. J. Training and proficiency of aviation electronics technicians: I. The proficiency of recent "A" school graduates (Bureau of Naval Personnel Technical Bulletin 60-10). San Diego: Naval Personnel Research Field Activity, September 1960. (a)

Jones, E. I., & Abrams, A. J. Training and proficiency of aviation structural mechanics: I. Proficiency of recent AMS school graduates (Bureau of Naval Personnel Technical Bulletin 60-10). San Diego: Naval Personnel Research Field Activity, September 1960. (b)

Laabs, G. J., Harris, H. T., & Pickering, E. J. A personnel readiness training program: Operation and maintenance of the 1200 PSI Steam Propulsion Plant (NPRDC Technical Report 77-36). San Diego: Navy Personnel Research and Development Center, August 1977.

Laabs, G. J., Panell, R. C., & Pickering, E. J. A personnel readiness training program: Maintenance of the missile test and readiness equipment (MTRE MK 7 MOD 2) (NPRDC Technical Report 77-19). San Diego: U. S. Navy Personnel Research and Development Center, March 1977.

Mackie, R. R. Research on factors influencing the interpretation of sonar signals (Final Report 776-6). Goleta, CA: Human Factors Research Incorporated for Office of Naval Research, June 1974.

Personnel qualification standard for damage control--qualification section 2, general damage control (NAVPERS 94119-2A). Washington, D. C.: Bureau of Naval Personnel, June 1971.

Personnel qualification standards (PQS) program (CNET Instruction 3500.3). Pensacola, FL: Chief of Naval Education and Training, June 1975.

Pickering, E. J., & Abrams, A. J. Experimental training of sonarmen in the use of electronic test equipment: III. The evaluation of the experimental program (Bureau of Naval Personnel Technical Bulletin 62-81). San Diego: Naval Personnel Research Activity, July 1962.

Propulsion Examining Boards for conventionally powered ships (OPNAVINST 3540.4B) Washington, DC: Chief of Naval Operations, November 1976.

PTEP Evaluation 5-76, Test Equipment Proficiency. Virginia Beach, VA: Central Test Site for the Personnel and Training Evaluation Program, Letter Report Serial 225 of 13 October 1976.

Rigney, J., Cremer, R., & Towne, D. The design of electronic equipment for ease of maintenance: Current engineering design practices (Technical Report 4). Los Angeles: University of Southern California, 1965.

Rigney, J., & Fromer, R. The Assessment of electronics corrective maintenance performance: I. Performance on the AN/URC-32 by Navy electronics Technicians. (Technical Report No. 42) Los Angeles: University of Southern California for Office of Naval Research, 1965.

Shriver, E. L., & Foley, J. P., Jr. Evaluating maintenance performance: The development of graphic symbolic substitutes for criterion referenced job task performance tests for electronic maintenance (AFHRL-TR-74-57 (III)). Wright-Patterson Air Force Base, Ohio: Air Force Human Resources Laboratory, November 1974.

Steinemann, J. H. Comparison of performance on analogus simulated and actual troubleshooting tasks (Personnel Research Activity Research Memorandum 67-1). San Diego: Naval Personnel Research Activity, July 1966.

Stuit, D., (Ed.) Personnel research and test development in the Bureau of Naval Personnel. Princeton, N.J.: Princeton University Press, 1947.

Training appraisal subsystem; establishment of curriculum and instructional standards (CIS) Offices/Departments (CNET Instruction 1540.6). Pensacola, FL: Chief of Naval Education and Training, July 1975.

Wilson, C. L., & Mackie, R. R. Research on the development of shipboard performance measures: Part I--The use of practical performance tests in the measurement of shipboard performance of enlisted naval personnel. Los Angeles: Management and Marketing Research Corporation for Office of Naval Research, November 1952.

Wilson, C. L., Mackie, R. R., & Buckner, D. N. Research on the development of shipboard performance measures: Part II--The use of performance rating scales in the measurement of shipboard performance of enlisted naval personnel. Los Angeles: Management and Marketing Research Corporation for Office of Naval Research, February 1954. (a)

Wilson, C. L., Mackie, R. R., & Buckner, D. N. Research on the development of shipboard performance measures: Part III--The use of performance check lists in the measurement of shipboard performance of enlisted naval personnel. Los Angeles: Management and Marketing Research Corporation for Office of Naval Research, February 1954. (b)

Wilson, C. L., Mackie, R. R., & Buckner, D. N. Research on the development of shipboard performance measures: Part IV--A comparison between rated and tested ability to do certain job tasks. Los Angeles: Management and Marketing Research Corporation for Office of Naval Research, February 1954. (c)

Winchell, J. D., Panell, R. C., & Pickering, E. J. A personnel readiness training program: Operation of the AN/BQR-20A (NPRDC Technical Report 77-4). San Diego: Navy Personnel Research and Development Center, December 1976.

ABOUT THE AUTHORS

Adolph V. Anderson received his undergraduate and graduate training at the University of Washington, Seattle, under such professors as Edwin Guthrie, Roger Loucks, Paul Horst, Allen Edwards, Ernest Hilgard, and Lloyd Humphreys. Since 1951 he has worked for the Navy in such areas of research and development as test development and validation, training, and performance evaluation.

Edward J. Pickering received his undergraduate training at San Diego State University and his graduate training at Wayne University, Detroit, Michigan. He came to the Navy Personnel Research and Development Center, then the Personnel Research Field Activity, in 1956. He is the author or coauthor of over 40 technical reports and papers. He has been involved in numerous research studies that have involved the utilization of job performance tests.

PERFORMANCE MEASUREMENT OF MAINTENANCE

John P. Foley, Jr.
Advanced Systems Division
Wright-Patterson Air Force Base, Ohio

ABSTRACT

This paper discusses the status of performance measurement (PM) for maintenance. During and after World War II both Navy and Air Force maintenance training programs made extensive use of formal job task performance tests. But for economy reasons, these tests were later abandoned in favor of paper and pencil theory and job knowledge tests. Considering the results of later research, these actions were most unfortunate. This research has indicated that such paper and pencil tests do not indicate how well individuals can perform the tasks of their jobs. Even though PM was used extensively during and after WW II, there have been few systematic research and development (R&D) efforts concerning the refinement of PM for maintenance. This paper describes briefly the AFHRL R&D efforts for PM which have given due consideration to the man-machine interface. The rather promising results of efforts to develop symbolic substitutes for PM are also presented. In addition, several problems concerning the research, development, and implementation of PM are discussed. The paper ends with proposals for future R&D efforts based on what has already been accomplished.

A LITTLE PERFORMANCE MEASUREMENT HISTORY

Performance Measurement (PM) efforts are not something new for the Defense Establishment. But many of the past PM efforts were not adequately documented. As a result, even their existence has been forgotten. From personal experience during World War II, I know that the training establishments of both the Army Air Force and the Navy made extensive use of such measurements for such maintenance job tasks as checkout, alignment, and troubleshooting.

I do not know exactly when PMs were de-emphasized in Navy maintenance training programs. However, in 1962 Harris and Mackie indicated that PMs were not being used in Navy training and field activities because it was generally felt that they required too much equipment and personnel time to be feasible.

In the Air Force, an active and substantial PM program continued until 1956. These measurement programs for the Air Training Command (ATC) of the Air Force and its predecessor, the Army Air Force, included elaborate checkrooms. To increase measurement objectivity and decrease instructor bias, these checkrooms were manned by full-time test administrators, whose sole job was to develop and administer both written and performance tests. In most cases, these checkrooms were equipped with their own hardware systems or subsystems which were used exclusively for PMs. This program required a substantial amount of equipment time, as well as test subject and test administrator time.

In 1956, almost all checkrooms were abolished to save money, equipment, and personnel. An often used argument in favor of this action was that most civilian schools did not have checkrooms and that instead, in civilian schools, the classroom, laboratory, or shop instructors were responsible for measurement and grading, which was true. But the weakness in this argument is that, in most cases, the shop instructors in civilian vocational schools did not have time to administer PM efforts and also supervise shop exercises. As a result, the Air Force had a far superior and more valid measurement system than civilian vocational schools. Another argument was that the resources required for PM efforts could not be justified since they were not part of the directed mission of the ATC.

But no matter what the reason, there was a drastic decrease in the number of PMs used in ATC after 1956. The decrease or elimination of PMs resulted in complete reliance on paper-and-pencil theory and job knowledge tests as measures of school success. The absence of PMs resulted in a decreased emphasis concerning "hands on" equipment exercises in maintenance training programs, especially those for electronic maintenance training.

Early Air Force R&D for Maintenance PM

Although the use of PMs in ATC did encourage the use of valuable "hands on" training, the PMs used did not reflect a systematic development process. As a result, their quality varied greatly from checkroom to checkroom. These and other weaknesses of the PMs used in ATC were recognized by personnel of the Maintenance Laboratory of the Air Force Personnel and Training Research Center (AFPTRC) in the early 1950s. (This Maintenance Laboratory, located at Lowry Air Force Base, was directed by Dr. Robert M. Gagne). This measurement research and development (R&D) continued until the demise of that laboratory in 1958.

One output of this effort intended for improvement of the development and administration of PMs was "A Guide for Use in Performance Testing in Air Force Technical Schools" (Highland, 1955). However, this useful document was published too late. Due to the closing of checkrooms and the resulting de-emphasis of PM, this guide received little or no use in ATC. However, if it had been followed, it certainly would have resulted in improved PM. One serious shortcoming of this guide, as viewed from today's vantage point, was the undue credence it gave paper-and-pencil measures.

In this regard, another important document of the Maintenance Laboratory reported the intercorrelations of measures concerning the proficiency of radar mechanics (Crowder, Morrison, and Demaree, 1954). This was one of the early studies which reported extremely low correlations between results of PM and results of paper-and-pencil theory and job knowledge tests. During the 1950s and early 1960s, there were a number of other studies that produced similar findings. This matter will be discussed later.

It certainly was unfortunate for the quality of maintenance that the use of PM was de-emphasized. But, at the time of these actions, much of the information now available about the weaknesses of paper-and-pencil tests for measuring school and job success had not been published. Even the most ardent supporters of checkrooms and PM in ATC had much more faith in the value of such paper-and-pencil tests than the subsequent research indicated. So, under such circumstances, one cannot be too critical of the decision makers who caused the elimination or de-emphasis of PM. Perhaps, if such information had been presented at that time, ATC would have retained their checkrooms and their PM.

Early PM Efforts of the Advanced Systems Division (AFHRL)

With the abolishment of AFPTRC in 1958 and the resultant closing of its Maintenance Laboratory, the Air Force maintenance research responsibility was transferred to the Behavioral Sciences Laboratory (BSL) at Wright-Patterson Air Force Base, Ohio, with greatly reduced manpower and monetary support. Since none of the research personnel was transferred with the responsibility and all of the ongoing projects had been cancelled, the research program, conducted by the Training Research Division of BSL, was not a true continuation of work of the Maintenance Laboratory. (In 1968, the Training Research Division of BSL became part of the newly formed Air Force Human Resources Laboratory and eventually was renamed the Advanced Systems Division (AS) of AFHRL).

The maintenance R&D supported by BSL and its successor, AFHRL/AS, has been characterized by its emphasis on the maintenance man's interface with the hardware being maintained, as well as the improvement of his efficiency of performance on the job. Before an extensive program was started, an in-house analysis was made concerning the variables that contribute to the performance of maintenance (see Foley, 1973, pp. 14-16). Eventually three closely related R&D programs resulted; namely, performance measurement, job performance aids (JPA), and job (task) oriented training (TOT). In each of these programs, a determined effort was made to make maximum use of the previous R&D conducted by Army, Navy, and Air Force including the AFPATRC work. The planning of new work for each program was preceded by an in-depth review and analysis of the R&D literature.

In regard to the literature reviews and analyses made for PM (Foley, 1967, and Foley, 1974), many valuable PM efforts have been reported by the Army, Navy, and Air Force. However, most of these efforts have not been systematic efforts, having as their prime objective the improvement of the state-of-the-art of PM. Rather, they have been ad hoc PM developments to support job-oriented training research programs. A notable exception was the work of the AFPTRC Maintenance Laboratory. (Another more recent systematic Army effort, accomplished by the Human Resources Research Organization (HumRRO), was not covered in these reviews (Vineburg and Taylor, 1972 a & b; Vineburg, Taylor, and Caylor, 1970; Vineburg, Taylor, and Sticht, 1970)). As to civilian R&D, during the initial PM literature review (Foley, 1967), a serious attempt was made to identify and include the results of PM R&D from the civilian vocational education establishment. None was found.

A substantial outcome of the review of other PM efforts was a consolidation of research results concerning the correlations between results of PMs for various maintenance tasks and paper-and-pencil theory tests, job knowledge tests, and school marks. As to their value for measuring ability to perform maintenance tasks, this research evidence gives a low rating to all of these paper-and-pencil based measures of school and job success. Table 1 shows correlations that have been obtained by comparing job task performance tests (JTPT) to theory and job-knowledge tests, both of which are paper-and-pencil tests. Table 1 also includes correlations of JTPT with school marks. As indicated earlier, school marks have been heavily weighted with the paper-and-pencil test scores. An examination of this table indicates that the correlations of JTPT scores with theory test scores are generally somewhat lower than with job-knowledge tests. None of these measures are sufficiently valid for use as substitutes for JTPT (Foley, 1967 and 1974).

THE MAN-MACHINE INTERFACE FOR MAINTENANCE

As stated previously, the maintenance R&D supported by AFHRL has emphasized the man-machine interface. From this point of view, PM for all personnel associated with machine systems must determine the ability of such personnel to perform tasks generated by the man-machine interface. Although there may be some overlap, most of the task functions demanded by a machine system of its operator personnel are different than those task functions demanded of its maintenance personnel. Herein, lies most of the unique, distinguishing characteristics of PM for maintenance. As a result, this section of my paper will be devoted to a discussion of the complexity of maintenance task functions.

Past Human Factors Emphasis

Before discussing the characteristics of task functions for maintenance, it might be well to call attention to the fact that human factors establishments have given much more attention to the operator interface with machines than they have given to the maintenance personnel interface. Many actions are taken to maximize effective and efficient performance of the operator. Work stations are human engineered to maximize the efficiency and comfort of the human operator. Major training facilities are provided, so that operators can receive a large amount of supervised practice in performing typical tasks of their job. Graduation from training is based primarily on demonstrated ability to perform job tasks. Further, periodic checks are made of the operator's ability to perform the critical tasks of his job. These, of course, are not all of the many efforts made to maximize the performance of human operators.

Generally, the human factors establishment has given little attention to the effectiveness and efficiency of the maintenance man's interface with hardware. The maintenance work, including the PM work of AFHRL/AS, has emphasized this neglected interface.

The Structure of the Man-Machine Interface for Maintenance

One of the results of our R&D for maintenance has been the evolution and articulation of a structure for handling maintenance functions and their complex relationships in a systematic manner. This structure includes: (1) standard maintenance functions and action verbs, (2) a working definition of a maintenance task, and (3) schemes for handling the complexities of maintenance tasks.

Standard Maintenance Functions and Action Verbs. The establishment of standard maintenance functions and action verbs has been one of the widely accepted results of the Air Force Systems Command's (AFSC) JPA effort entitled "Presentation of Information for Maintenance and Operation" (PIMO). (Although the PIMO project was managed by the Space and Missile Systems Organization (SAMSO) of AFSC, AS provided active participation and technical inputs during the entire project from 1966 through 1969. AS has incorporated the key findings and outputs of PIMO in its own JPA efforts.) Early in the PIMO project, it was found that many maintenance action verbs and functions were used by maintenance people, some with several different meanings. Part of this confusion was caused by the language used in maintenance technical orders, which were written by different people and produced by many different hardware manufacturers. As a result, maintenance technicians themselves did not generally use precise language. A

study was made to identify and define these action verbs. Where two or more verbs were used to indicate a similar action, the preferred verb was selected, based on the expressed preferences of a sample of maintenance men with a wide range of maintenance AFSC. The use of the preferred verbs of this list is now a firm requirement of Air Force technical order specifications, as well as of recent Army and Navy specifications (see Joyce, Chenzoff, Mulligan, and Mallory, 1973, pp. 97-142).

A Working Definition of a Maintenance Task. Within this list of action verbs are a number of key action verbs (functions). A key action verb, with an appropriate specific hardware unit as its predicate, becomes a task statement. Such a task statement represents a maintenance task which can be demanded by the existence and operation of a specific machine subsystem. A list of these functions is found in AFHRL-TR-73-43(I) (Joyce et al., 1973, pp. 19, 20). This list includes functions which are found in both mechanical and electronic jobs. Some apply to only mechanical jobs and some apply to both.

Schemes for the Systematic Consideration of Maintenance Functions and Tasks. Three schemes have been developed for the systematic consideration of maintenace functions and tasks and the key factors that affect them.

1. Scheme One. A convenient model for categorizing these maintenance functions with relation to the type of hardware and the level of maintenance is presented in Figure 1. The common maintenance functions, already mentioned together with the usage of test equipment and hand tools, are represented on one axis of the model. Since mechanical and electronic subsystems usually require a different variety of maintenance actions, they are represented by another axis. (In regard to this axis, mechanical maintenance could be further divided into two categories, one represented by hardware such as jet engines; and another, by hardware such as airframes and tank and ship hulls.)

The third axis of the model represents the three levels or categories of maintenance now found in the military services. Organizational maintenance, the first level, is usually aimed at checking out a whole machine subsystem and correcting any identified faults as quickly as possible. Flight line maintenance falls in this category. If a system is checked out and it does not work, the line replaceable unit (LRU) or "black box" causing the malfunction is identified and replaced. This major component is then taken to the field shop (intermediate maintenance) where it is again checked out and the faults, authorized for correction, are corrected. The corrective actions, authorized at the intermediate level, vary greatly from system to system depending on the maintenance concept of each system. On some systems, the maintenance man will troubleshoot the "black box" to the piece part level. In more modern equipment, he will identify a replaceable module made up of many piece parts. Some modules are thrown away, others sent to the depot for repair. Any LRUs which the field shops are unable, or unauthorized, to repair are sent to the depot for overhaul.

Organizational and intermediate level organizations are manned primarily by enlisted technicians whose average length of service is rather short (slightly more than 4 years in the Air Force). Depots are manned largely by civilian personnel with a much higher level of experience and longer retention time. Using this model, it has been possible to specify areas of concentration for study.

Since PM requirements for maintenance are so different for the various blocks indicated in this model, it is extremely important that PM researchers indicate the precise blocks of their concentration. To date, the AFHRL/AS has concentrated on the shaded electronic portions of this model (Figure 1). The resultant model battery of 48 JTPT, together with their symbolic substitutes, will be described later. In addition, a battery of 11 JTPTs was developed on an ad hoc basis (Shriver and Foley, 1975) for mechanical tasks at the organizational level of maintenance (see shaded portion of Figure 2). The HumRRO work, mentioned previously (Vineberg and Taylor, 1972 a & b; Vineburg, Taylor, and Caylor, 1970; Vineburg, Taylor, and Sticht, 1970) was concerned with mechanical hardware (tank and truck). The 13 tests developed concerned the maintenance functions which are indicated by the shaded portions of Figure 3.

2. Scheme Two. Maintenance functions have limited meaning unless applied to specific hardware. A task identification matrix (TIM) is an extremely effective and necessary device for interfacing these maintenance functions with the appropriate hardware units and thus identifying the maintenance tasks that are generated by a specific machine subsystem (see Figure 4). The TIM, when properly structured, will reflect the maintenance level or levels of interest; that is, organizational, intermediate, and/or depot. AFHRL-TR-73-43 (I) (Joyce et al., 1973, pp. 16-37) provides detailed directions for developing a TIM.

3. Scheme Three. A matter of serious concern when developing and structuring PM for maintenance tasks is the interaction among the maintenance tasks for one hardware. A four-level hierarchy of dependencies can be stated. Figure 5 gives a graphic presentation of these dependencies among maintenance activities for an electronic hardware.

The checkout of the AN/APN-147 (Doppler Radar), for example, can be a task in its own right. But the same checkout activity becomes an element of other major tasks such as calibrate. The calibration of doppler radar includes the operation of specific general and special test equipments, the use of specific hand tools, as well as the checkout activity. Troubleshooting of an electronic equipment, such as AN/APN-147, requires the use of general and special test equipments. It may be necessary to remove and install activities and/or to adjust, align, and calibrate activities. Efficient troubleshooting practice usually requires the use of a cognitive strategy to adequately track the dependent activities (but the cognitive strategy in itself is not troubleshooting). Any troubleshooting task should begin and end with an equipment checkout. Because of these various and varying dependency relationships, such activities as checkout, remove, install, disassemble, adjust, align, calibrate, or troubleshoot cannot legitimately be considered as discrete tasks, even for one electronic system.

Another confounding factor is the false correspondence that the same functional verbs create when applied to different electronic hardware. For example, personnel with the Avionics Inertial and Radar Navigation Specialist, AFSC 328X4, are maintaining at least 50 major electronic subsystems. Many vintages of hardware design are represented. The checkout activity for each is different (both in content and difficulty) and, in some cases, very different. The lack of correspondence of alignment, calibration, and troubleshooting tasks from one specific equipment to another is even greater. An example of the lack of correspondence from one hardware to another is the wide difference in the content and difficulty of troubleshooting tasks between two doppler radars. The AN/APN-147, which is

used on the C-130 and C-141, has approximately 14,000 shop replaceable units (SRU) whereas the IDNE on the C-5 has only 28. This lack of correspondence of functions across electronic hardware makes it difficult to generalize from results of PM from one electronic hardware to another. One exception is in the area of general test equipment which may be used in performing maintenance tasks across many hardware subsystems.

The examples given are characteristic of many of the electronic maintenance AFSC. Similar problems in complexity of maintenance functions and tasks are found in mechanical hardware, but to a lesser degree.

Maintenance Functions and Tasks and Traditional Psychological Variables

In this consideration of the characteristics of maintenance functions and tasks, the psychological language normally used by human factors specialists in describing the activities of operator personnel has not been used. There are several reasons for this. Such analyses would be extremely expensive to generate and would be of little value to maintenance, personnel, and training people. In most cases, a task (generated by a maintenance functional verb plus its specific hardware unit) is considerably different from another task (generated by the same functional verb plus a different hardware unit). A separate human factors analysis would have to be made for each of these tasks. Some maintenance specialities now include over 50 major electronic subsystems--most of which produce hundreds of such tasks.

A traditional human factors type of task analysis for such tasks, if properly utilized, would probably be of great value during the original design of a specific hardware or for the design of realistic training simulators. But most of maintenance personnel interface with such subsystems long after their design. The type of task analysis required for the maintenance man calls for a different language. The functions used in this discussion of PM are, therefore, based on a common language that is familiar to (if not always completely understood by) a wide range of DOD personnel directly or indirectly associated with maintenance.

DEVELOPMENT OF PM AND SYMBOLIC SUBSTITUTES FOR PM

Starting in 1969, the Advanced Systems Division of the AFHRL supported a modest program to provide the Air Force with the necessary tools for measuring the ability of maintenance personnel to perform the key tasks of their jobs. The scope of this work was limited to the maintenance of electronic hardware at the organizational and intermediate levels (see shaded portion of Figure 1). This program had two objectives: (1) to develop a model battery of JTPTs together with appropriate scoring schemes for the measurement of the task performance ability of electronic maintenance personnel (an effort was to be made for the development of JTPTs which could be easily administered), and (2) using these JTPTs as criteria, to develop and try out a series of paper-and-pencil symbolic substitute tests that would hopefully have high empirical validity.

Criterion Referenced Job Task Performance Tests

A model battery of 48 criterion-referenced JTPTs and a test administrator's handbook were developed for measuring ability to perform electronic maintenance

tasks. Copies of the actual instructions for test subjects together with the test administrator's handbook are available from the Defense Documentation Center (DDC) as AFHRL-TR-74-57 (II), Part II (Shriver, Hayes, and Hufhand, 1975). The test administrator's handbook was developed with step-by-step detailed instructions so that an individual with a minimum of electronic maintenance experience can administer the tests.

The battery includes separate tests for the following classes of job activities: (1) equipment checkout, (2) alignment/calibration, (3) removal/replacement, (4) soldering, (5) use of general and special test equipment, and (6) troubleshooting. The Doppler Radar AN/APN-147 and its Computer AN/ASN-35 were selected as a typical electronic system for use as the testbed for this model battery. The soldering and general test equipment JTPTs are applicable to all electronic technicians. The other tests of the battery apply to technicians concerned with this specific doppler radar system. A detailed description of the development and tryout of these JTPTs is given in AFHRL-TR-74-57 (II) (Shriver and Foley, 1974a). Each class of activity for which JTPTs were developed contains its individual mix of behaviors, but it is not mutually exclusive. As indicated in Figure 5 and Table 1, a four-level hierarchy of dependencies exists among them.

After considering product, process, and time as to their appropriateness for scoring the results for each activity, it was decided that a test subject has not reached criterion until he had produced a complete, satisfactory product. This was a go, no-go criterion.

Table 2 summarizes the number of tests, problems, and scorable products by class developed for the AN/APN-147 and AN/ASN-35. The simple addition of numbers shown in Table 2 indicates that there are 48 tests, 81 problems, and 133 scorable products. But, these numbers tell us nothing in terms of the content of the tests. To say that one test subject accomplished 100 scorable products while another accomplished 90 tells us nothing about the job readiness of these individuals or that one is better than the other. The varieties of scorable products are so diverse that any combination of them, without regard to what they represent, is meaningless. The only meaningful presentation of such information must be in terms of a profile designed to attach meaning to such numbers. A sample of such a profile is shown in Figure 6.

This profile is not presented as the final solution to the profile problem for JTPTs for electronic maintenance. It does contain most of the important information regarding a test subject's success on the full range of tests. Also, it gives a meaningful picture of the subject's job task abilities as measured by the test battery, indicating the subject's strengths and weaknesses.

An examination of the profile (Figure 6) indicates that most of the tests in this battery contain only one problem. For example, there are two checkout tests, having one problem each and there are 11 troubleshooting tests having one problem each. There are two soldering tests; one has two problems and the other has three. The voltohmmeter (VOM) test has 20 problems.

The subject receives no "credit" for a problem unless he obtains all of the expected products. No attempt is made to combine these scores in terms of meaningless numbers.

The hierarchy of dependencies discussed previously (Figure 5) has implications for the order in which tests are administered, as well as for diagnostics. For example, since troubleshooting includes the use of test equipment and other activities in the hierarchy, logic would dictate that, in most training situations, the administration of the tests for the subactivities would precede the troubleshooting tests and that a test subject would not be permitted to take the troubleshooting tests until he had passed these other subtests. Under some circumstances, one may wish to reverse the process. A subject who successfully completes selected troubleshooting or alignment tests can be assumed to be proficient in his use of test equipment and checkout procedures. These dependencies are displayed on the left-hand side of the profile (Figure 6).

Due to the unavailability of a sufficient number of experienced test subjects at the time of the tryout of the JTPT battery, the tryout was not as extensive as planned. The limited tryout did indicate that the tests as developed are administratively feasible. Their continued use, no doubt, would result in further modifications and improvements.

Development of Symbolic Substitutes

There is no doubt that a battery of JTPT would require more training and on-the-job time of the test subjects, more equipment, and specially trained test administrators. It will be recalled that these requirements were high among the reasons given for dropping PM from the Air Force and Navy maintenance training programs. Therefore, the availability of empirically valid symbolic substitute tests would be highly desirable. Even though previous attempts to develop such tests as the Tab Test (Crowder et al., 1954) had failed, it was our opinion that much more work could be done to improve symbolic maintenance tests as substitutes for JTPT. It was hypothesized that higher correlations possibly could be obtained by a different approach to the development of symbolic tests. A study of the Tab Tests (Crowder et al., 1954, see Table 1) indicated that the JTPTs used as the criterion measures contained many distractions and interruptions to the subject's troubleshooting strategy (cognitive process), such as using test equipment to obtain test point information. In addition to such interruptions to the cognitive process, the subject can obtain faulty test point information by the improper use of his test equipment. In the symbolic substitute Tab Tests, all of these potential pitfalls of the actual task were avoided. The subject was given a printed test point readout. It was hypothesized that the injection of job equivalent pitfalls into symbolic substitutes possibly would increase their empirical validity.

Based on these hypotheses, a battery of symbolic tests was developed under contract with the Matrix Research Company of Falls Church, Virginia. A companion graphic symbolic test was developed for each of the job activities for which a criterion-referenced JTPT had previously been developed. Based on two limited validations, all of the graphic symbolic tests, with exception of the symbolic test for soldering, indicated sufficient promise to justify further consideration and refinement. Table 3 indicates the correlations obtained from these validations. Due to a shortage of available subjects, the number of pairs of subjects was extremely small. All of these promising graphic symbolic tests, therefore, must be given more extensive validations using larger numbers of experienced subjects.

The validation of any such symbolic test requires the administration of a companion JTPT as a validation criterion. As a result, a validation is an expensive process in terms of equipment and experienced manpower. The troubleshooting symbolic tests require the most extensive refinement. Several suggestions are made for improving their empirical validity. A complete description of these symbolic test efforts can be found in AFHRL-TR-74-57 (III) (Shriver and Foley, 1974b). Also, an attempt was made to develop video symbolic substitute tests, but this effort produced no promising results (Shriver, Hayes, and Hufhand, 1974).

Even if graphic symbolic substitutes of high empirical validity can be produced, the use of symbolic substitutes will never, in my opinion, dispense with the requirement for the liberal administration of actual JTPT to maintenance personnel. We can never include all aspects of an actual performance of a task in a paper-and-pencil symbolic representation of that task, but our work indicates that we can come much closer than has been done in the past.

CONSOLIDATED DATA BASE TO SUPPORT PM

In keeping with its man-machine interface orientation, AFHRL/AS is demonstrating the technical feasibility of integrating five human resources related technologies and applying them during weapon system development. This is being accomplished under Project 1959, "Advanced System for the Human Resources Support of Weapon System Development."

The five technologies are:

1. Human Resources in Design Tradeoffs
2. Maintenance Manpower Modeling
3. Job Performance Aids
4. Instructional System Design
5. System Ownership Costing

One objective of this program is to determine the data input requirements for, and prepare specifications for, a consolidated maintenance task identification and analysis data base that will support the integrated application of these five technologies in a weapon system development program. We feel that such a consolidated data base will contain most, if not all, of the information which would be required to develop good JTPT, provided the tests are developed in keeping with the technology described in this paper. If such a data base is demonstrated to be technically feasible and if it is routinely made a requirement in weapon system development contracts, it will provide considerable assistance in developing maintenance performance tests for new weapon systems.

PROBLEMS CONCERNING THE RESEARCH, DEVELOPMENT AND IMPLEMENTATION OF PM

As stated previously, PM for maintenance had widespread usage in Air Force and Navy maintenance training programs during and after World War II. The dropping of such tests from these training programs reflected two interacting prime factors. The first prime factor is a fact, that is, PM tests for maintenance

are much more expensive to develop and to administer than paper-and-pencil theory and job knowledge tests. However, the second factor, the general acceptance of such tests as adequate substitutes for PM, is not a fact but a widely held belief. I use belief here with the precise meaning of something that is held to be true without adequate proof. Although we now have substantial hard data which disproves this belief (see Table 1), many people seem to be unaware of these data. Most of the objections to PM ignore the fact that paper-and-pencil tests are not valid measures of job ability. Such paper-and-pencil tests are not a bargain. No matter how cheaply they can be administered, their results are almost meaningless in terms of measuring ability to perform maintenance tasks. This state of affairs has contributed to a number of other problems.

1. There is a well developed paper-and-pencil test technology which is based on testing theory which is appropriate for the academic variety of education. This technology has been institutionalized and is well entrenched in the DoD personnel and training systems. All education test and measurement text books and courses reflect this technology. Psychological measurement texts emphasize this technology. At least two generations of teachers and test and measurement psychologists have been trained in the use of this technology and, as a result, many have unquestioned faith in its application to any personnel measurement problem.

Most of these people are products of the academic world. Few have had any "hands on" experience in performing maintenance tasks. When the appropriateness of their technology for the measurement of maintenance ability is questioned, many members of this paper-and-pencil testing establishment become threatened and, therefore, defensive.

2. In spite of this extensive military history of usage, there is no PM establishment comparable to the paper-and-pencil test establishment. There are no college test and measurement courses (even in vocational education departments) which teach PM technology, and there are no text books devoted to the subject. The vocational educators have emulated their academic brethren by using their measurement texts. There has only been a limited amount of systematic R&D concerning the development of a PM technology. Most of the current PM technology for maintenance is found in DoD technical reports.

3. Just as human factors resources have favored the operator's interface with hardware over that of the maintenance man's interface, the personnel and training resources have heavily favored the operator. This has been especially true with regard to the aircraft pilot. DoD still contains elements of a caste system which relegated the maintenance man to the status of a "grease monkey." There is a reflection of a deep-seated culture bias in our society against any group who gets their hands dirty while earning their living. This bias has been extremely strong in the management and academic establishments. The importance of the maintenance man and his problems has been consistently downgraded, perhaps not by word, but certainly by the allocations of resources. No matter how costly, the operator has always been provided the necessary hardware and hardware simulators, as well as the necessary PM, to ensure his ability to perform the tasks of his job. Few such facilities have been provided for the maintenance function--one result has been an effective but inefficient and costly maintenance system. Costly maintenance is directly translated into excessive life cycle costs of ownership of hardware.

4. Success in aircraft pilot training and other operator training has always been based on PM, that is, demonstrated ability to perform key job tasks. Consequently, these training programs have been designed to ensure success on PM and have been characterized by an abundance of supervised practice of job tasks. But, for maintenance personnel, paper-and-pencil theory and job knowledge tests have been used as the principal means for determining both the school and job success. As a result, maintenance training programs, both formal courses and career development courses (CDC), have come to be structured to ensure success on paper-and-pencil tests. This has resulted in the greater part of many so-called maintenance courses taking on the verbal characteristics of academic education. This has happened at the expense of supervised practice of job tasks.

5. A like imbalance of emphasis is reflected in the more stringent PM certification required of the operator. A pilot, for example, is certified on the basis of his demonstrated performance before he is permitted to fly a specific type of aircraft, and his proficiency is checked periodically as long as he is required to fly that aircraft. But a maintenance man receives no such certification of his ability to perform the maintenance tasks required of him by the same aircraft.

Rather than an equipment specific PM certification, an "occupational" certification based on paper-and-pencil job knowledge tests has been substituted for maintenance personnel. Many maintenance "occupations" cover a large number of systems or subsystems. An individual maintenance man usually works on one or two such systems or subsystems. Tests for occupations have, therefore, been general in nature. Most of the personnel and training measures for maintenance men in all three services have been of the paper-and-pencil job knowledge variety. However, the Army now has a policy for including PM on specific job tasks in its maintenance personnel system (Maier, Young, and Hirshfeld, 1976). This policy is only in an early stage of implementation.

Returning to the pilot/maintenance man comparison, it is true that an improficient pilot might destroy a whole aircraft. Thanks to good checkout procedures, it is highly improbable that a maintenance man's actions would cause the sudden destruction of a whole aircraft. However, over a period of time, an improficient maintenance man can do the equivalent, on a piece-by-piece basis, by the damage he can cause by his lack of skill and by his consumption of unnecessary spare parts to correct malfunctions. Certification by PM would certainly improve the efficiency of maintenance.

6. Closely related to this lack of meaningful certification for maintenance is the lack of accountability. The target of the personnel, training, and tech data establishments should be to ensure the maintenance man's ability to perform the tasks of his job efficiently. But our personnel measures do not ascertain how many hits and misses we make--nor what is causing our misses. As a result, no one is being held accountable for the effectiveness of their contributions in terms of efficiency of job task performance. Many people in these establishments can see no reason for adopting improved technologies such as TOT and JPA --because they have never been held accountable for hitting the job performance target. We, therefore, require the use of valid job task performance measures to provide the bases for such required accountability. But such a possibility becomes very threatening to many people in these establishments.

7. In spite of all of the evidence supporting requirements of PM for maintenance, it has been extremely difficult to obtain R&D funding for efforts to advance the PM technology. In addition, difficulty has been experienced in finding and retaining Air Force professionals with the necessary capability and interest to pursue an effective PM R&D program for maintenance. Such professionals are necessary, either for an in-house or contractor program.

Few contractors have had extensive experience or expertise in this area. Any contractual effort, to be effective, must be very carefully planned and closely monitored. I would anticipate that much of the first year's effort by a new contractor will be expended in a learning experience for his people and will not be too productive for the PM technology. Unless continued follow-on work is given such a contractor, his expertise is soon lost.

During Fiscal Years 1969, 1970, and 1971, a total of $239K in exploratory development funds was obtained by AFHRL/AS for the development and tryout of PM and symbolic substitutes. The contractor personnel for this effort developed considerable expertise in working with PM for maintenance but they are no longer with the original contractor. The principal investigator, Dr. Edgar L. Shriver, is now president of his own firm, but his two PM assistants are no longer with him. Any successful program in this PM area must be a long-range program making use of existing expertise and aimed at expanding such expertise. "Off again, on again" efforts, jumping to a new contractor with every start, will result in little improvement in PM technology.

8. During JTPT and symbolic test development efforts, several attempts were made to share the use of operational hardware on a noninterference basis. These experiences have indicated that, no matter how cooperative the personnel of the operational unit, such time-sharing efforts are very expensive in terms of wasted man-hours of highly paid R&D professional personnel. For successful results, the necessary hardware must be assigned to the R&D project.

9. One of the persistent problems concerning the administration of PM has been getting maintenance supervisors to shed their supervisory role and assume the role of a disinterested test administrator. Because of their strong urge to show and help test subjects, most of these people have extreme difficulty in keeping themselves out of the actual task performance.

10. Timewise, it certainly would be impossible to administer a PM to a maintenance man for every possible task that his hardware system might produce. This world of tasks and people must be sampled. The model PM described previously provides a sampling procedure based on major task functions such as checkout, align, adjust, troubleshoot, etc. But even this sampling across possible tasks resulted in 48 tests and 133 scorable products. It would be impractical to give any one test subject all of these 48 tests at any one time. Systematic sampling schemes must be developed across tests.

The purposes for which PM results are to be used should be considered when developing sampling schemes. Such purposes of PM could include ascertaining: (1) the job task proficiency of an individual, (2) the job effectiveness of a training program, and (3) the proficiency of a maintenance unit. Each of these purposes would require a different mix or mixes of tests and people. Some suggestions for such samplings can be found in AFHRL-TR-74-57 (II), Part I (Shriver and Foley, 1974a). But it should be remembered that these are suggestions that must still be field tested.

In the case of determining unit proficiency, some PM can be administered by on-line observation of tasks that are often repeated such as checkout. There will always be a requirement for off-line PM concerning critical, but seldom performed tasks. Whether the PM is performed on-line or off-line, the test administrator must use the same objective scoring procedures, the criteria of success being an acceptable product.

11. The potential cost of PM in both training and field environments has certainly been increased by the proliferation of numbers of hardware subsystems (especially electronic) since the early 1960s. Over this period, the state of the art has been constantly changing. This has resulted in the proliferation of many variations in tasks for any one task function. For example, the alignment function produces considerably different tasks from hardware to hardware. Some long-range actions are being taken to reduce the number of hardware having the same functional use. Because of the large numbers and types of maintenance tasks, a realistic system of priorities must be established for PM development. PM concerning the use of general test equipment would probably have the most immediate and widespread effect on the quality of maintenance. This development should be followed by PM for systems and subsystems having long life expectancies and large numbers in the field.

12. Current military grading systems must be modified to properly reflect the results obtained from PM and symbolic substitutes. In my opinion, the only adequate device for presenting such results is a profile similar to that shown in Figure 6. No attempt should be made to convert the content of such a profile into a single numerical score. The results of PM should never be combined with paper-and-pencil test results.

Institutionalization of New Technologies

Getting newly developed technologies such as PM institutionalized is a perennial problem, especially when a technology requires fundamental changes in long existing programs, procedures, and attitudes of entrenched establishments. AFHRL/AS has been involved in the implementation of several well developed and documented technologies such as job performance aids and instructional systems design (ISD), including programmed instruction and job (task) oriented training. These experiences have indicated that it is extremely difficult to maintain the integrity of a technology during its so-called implementation. Operational organizations invariably attempt to implement a much "watered down" version of the technology and consequently obtain much "watered down" results. In some cases, only cosmetic changes to existing programs are reported as implementations. Currently it requires years of persistent effort on the part of the research community to get a technology properly institutionalized.

A mechanism must be developed for the timely institutionalization of each new technology which will ensure its integrity. A mechanism for the orderly implementation of technologies similar to that used for new weapons systems is recommended. Such a mechanism must make efficient and effective use of the "know how" of the developers of the technology and make them responsible and accountable for its implementation. A new technology should not be turned over to a using command for its operation until it is in place, "debugged," and operational--just as a new weapons system is not turned over to an operational command until it has been "debugged" and proven to be ready for operational use.

PROPOSED PM R&D EFFORTS FOR MAINTENANCE

Excessive maintenance costs are never going to be reduced as long as we don't have JTPTs and/or empirically valid symbolic substitutes to ascertain how efficiently maintenance men perform the tasks of their jobs. In my opinion, the lack of such measures of maintenance performance is a most serious deficiency in DoD. As such, R&D in this area should have an extremely high priority.

Areas for R&D Concentration

For a long-range R&D effort, five general areas of concentration are recommended; namely, JTPT and matching symbolic substitute tests for electronic maintenance, JTPT and matching symbolic substitute tests for mechanical maintenance, and aptitude tests based on PM. The development and field tryout of a JTPT must precede the development of its symbolic substitute. The work on JTPT batteries for both electronic and mechanical maintenance should be started as soon as possible. The work on aptitude tests should not be started until JTPT batteries and the symbolic substitute tests have been completely field tested. More information concerning these areas of concentration follows:

1. Refinement of Model JTPT Battery (Electronic Maintenance). The already available model JTPT Battery (Shriver et al., 1975) should be given a large scale field tryout. (Since the AB328X4 Avonics Inertial and Radar Systems Specialist Course, which includes the AN/APN-147 and the AN/AJN-35, does not emphasize the mastery of job tasks, the equipment specific tests of this battery cannot be used in the formal course.) One thrust of this effort should be to further refine the battery including its administrative procedures. A second thrust should be the development of sampling strategies which would be appropriate for determining the effectiveness of training programs and both individual and unit proficiency as discussed earlier under PM problems. This effort would require approximately 2 professional man-years plus the use of maintenance specialists as test administrators from the appropriate maintenance specialties. If it is necessary to select a system other than the AN/APN-147-AN/AJN-35 combination, this work would require approximately 4 professional man-years.

2. Refinement of Symbolic Substitutes (Electronic Maintenance). As previously indicated, several symbolic substitutes for JTPT were developed and given a limited tryout. Although Table 3 indicated that some of the symbolic tests show promising empirical validity, they must be more thoroughly refined and validated. In addition, further exploratory development is required for symbolic substitute tests for troubleshooting tasks in keeping with recommendations made in AFHRL-TR-74-57 (III) (Shriver and Foley, 1974b). This effort would require between 3 and 4 professional man-years plus the use of maintenance specialists as test administrators and test subjects from the appropriate maintenance specialties.

3. Development of Model JTPT Battery (Mechanical Maintenance). A model JTPT battery similar to the model battery for electronic maintenance described previously should be developed for a typical mechanical subsystem such as a jet engine or tank engine. This model should cover both the organizational and intermediate levels of maintenance, and should be thoroughly field tested. Sampling strategies as indicated for the electronic battery should also be developed. This effort will require approximately 4 professional man-years plus the use of maintenance men from the appropriate maintenance specialities as test administrators and test subjects.

4. Development of Symbolic Substitutes (Mechanical Maintenance). An attempt should be made to develop symbolic substitute tests with high empirical validity after the model JTPT battery is available for mechanical maintenance. The same contractor that developed the JTPT battery should develop these symbolics. A very rough estimate for accomplishing this symbolic effort would be 4 professional man-years.

5. Job Aptitude Test Research Based on Results on JTPT. R&D plans should be made to utilize the results of JTPT and symbolic substitute tests for standardizing military aptitude indices obtained from the Armed Service Vocational Aptitude Battery (ASVAB). As a first step, the military aptitude scores of all tests subjects used for the tryouts in the proposed JTPT R&D should be recorded. In addition, such aptitude scores should be obtained during any school or field administration of JTPT or symbolic substitutes. When sufficient data are obtained, the degree of relationship between JTPT results and various aptitude indices should be obtained. Later, when a sufficient number of JTPT are used in the field, a formal R&D project should be initiated to modify the ASVAB to directly reflect job success as measured by JTPT.

R&D Strategy

Probably the most cost-effective approach for PM for both electronic and mechanical maintenance would be to concentrate on the development and refinement of JTPT on use of key test equipments prior to proceeding with the other task functions of the proposed model test batteries. As indicated in Figure 5, the use of general test equipment is a prerequisite to maintenance task functions such as alignment, calibration, and troubleshooting. In addition, general test equipments usually have wide usage in such task functions across many hardware systems and there is a substantial amount of data which indicates that many maintenance men are weak in their test equipment ability. So, a general improvement in ability to use test equipment is an important and necessary factor for the general improvement of several maintenance task functions. I would strongly recommend, therefore, that the early concentration for the proposed model test batteries be in JTPT concerning the use of key test equipments. Each PM development for a test equipment should be accompanied by the development of a programmed training package with sufficient practice frames for teaching the mastery of all its functions. Basic models of such training packages for 12 general test equipments are now available (see Scott and Joyce, 1975a through l). However, more practice frames should be included in these programs.

Table 1 Correlations Between Job-Task Performance Tests and Theory Tests, Job Knowledge Tests, and School Marks

Researchers	Type of Job Task Performance Test (JTPT)	Theory Tests	Job Knowledge Tests	School Marks
Anderson (1962a)	Test Equipment JTPT			.18 - .33
Evans and Smith (1953)	Troubleshooting JTPT	.24 & .36	12 & .10	.35
Mackie et al., (1953)	Troubleshooting JTPT	.38		.39
Saupe (1955)	Troubleshooting JTPT		.55	.56
Brown et al., (1959)	Troubleshooting JTPT	.40		
	Test Equipment JTPT		.29	
	Alignment JTPT		.28	
	Repair Skills JTPT		.19	
Williams and Whitmore (1959)	Troubleshooting JTPT			
	(Inexperienced Subjects)	.23		
	(Experienced Subjects)	.15		
	Adjustment JTPT			
	(Inexperienced Subjects)	.07		
	(Experienced Subjects)	.21		
	Acquisition Radar JTPT			
	(Inexperienced Subjects)	.03	.36	
	(Experienced Subjects)	.14	.22	
	Target Tracking Radar JTPT			
	(Inexperienced Subjects)	.24	.33	
	(Experienced Subjects)	.20	.38	
	Missile Tracking Radar JTPT			
	(Inexperienced Subjects)	.09	.15	
	(Experienced Subjects)	.19	.32	
	Computer JTPT			
	(Inexperienced Subjects)	.08	.24	
	(Experienced Subjects)	.06	.14	
	Total JTPT			
	(Inexperienced Subjects)	.14		
	(Experienced Subjects)	.20		
Crowder et al. (1954)	Troubleshooting JTPT	.11	.18 - .32	

Table 2 Tests, Problems, and Scorable Products

Class	Code	Tests	Problems	Scorable Products
1. Checkout	CO	2	2	2
2. Physical Skill Tasks (soldering)	PT	2	5	17
3. Remove and Replace	RR	10	10	20
4. Test Equipment	SE	7	37	67
5. Adjustment	AD	6	6	6
6. Alignment	AL	10	10	10
7. Troubleshooting	TS	11	11	11
Total	7	48	81	133

Table 3 Indicates the Number of Pairs Used as Well as the X^2 and the Correlations Obtained during Two Small Validations of Symbolic Tests

Test Area	N Pairs	x^2	ϕ	r_t
Novice Subjects (Altus)				
Checkout	4	4.00	1.00	–
Remove & Replace	14	2.57	.43	–
Soldering Tests	4	0	0	–
General Test Equip	6	2.67	.67	–
Special Test Equip	6	.67	.33	–
Alignment/Adjustment	19	6.37	.58	–
Troubleshooting	9	1.00	−.33[a]	–
Experienced Subjects (TAC)				
Overall Troubleshooting	30	6.53	.47	.68
Chassis (Black box)				
Isolation	30	16.33	.73	.81
Stage Isolation	30	3.33	.33	.46
Piece/Part Isolation	15	.07	.07	.16

[a] This negative correlation was probably due to a number of deficiencies such as (1) deficiencies in the Fully Proceduralized Job Performance Aids provided the subjects, (2) deficiencies in the sequencing of the troubleshooting JTPT in relation to the sub tests in the JTPT battery, (3) maintenance difficulties with the AN/APN-147 AN/ASN 35 system, and (4) difficulties with the content and administration of test equipment pictorials provided in the original troubleshooting symbolic tests.

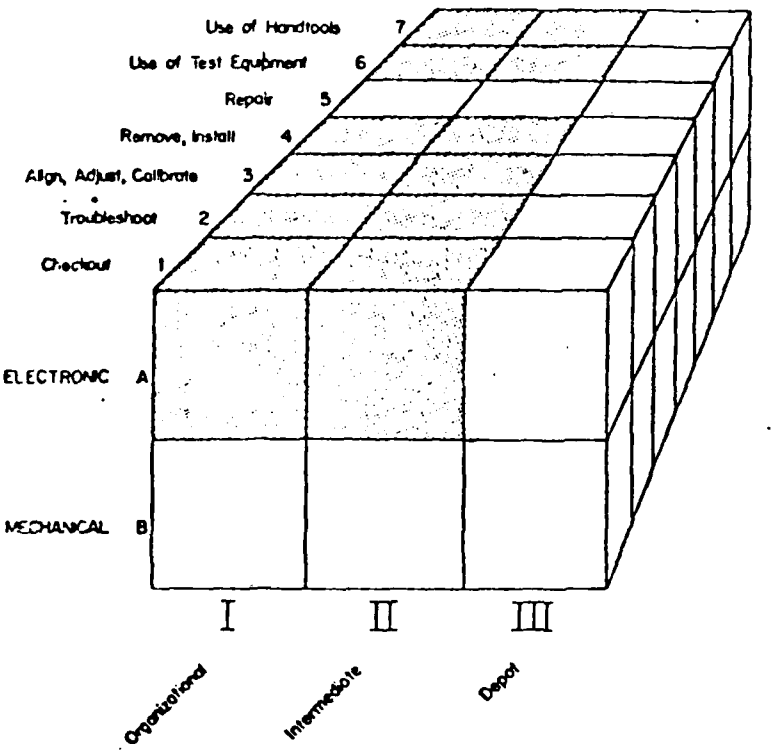


Figure 1 - A functional representation of the DOD Maintenance Structure (Shaded portion indicates scope of AFHRL PM development for electronic maintenance).

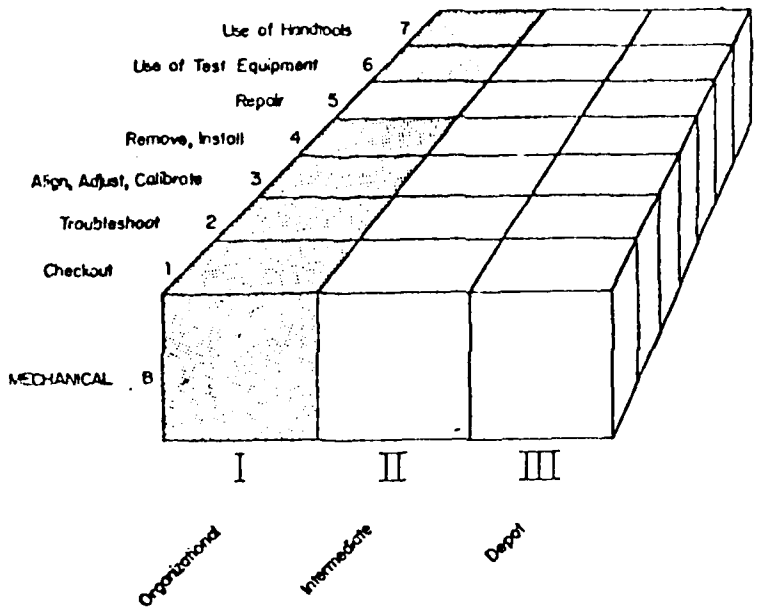


Figure 2 - A functional representation of the scope of AFHRL FM development for mechanical maintenance.

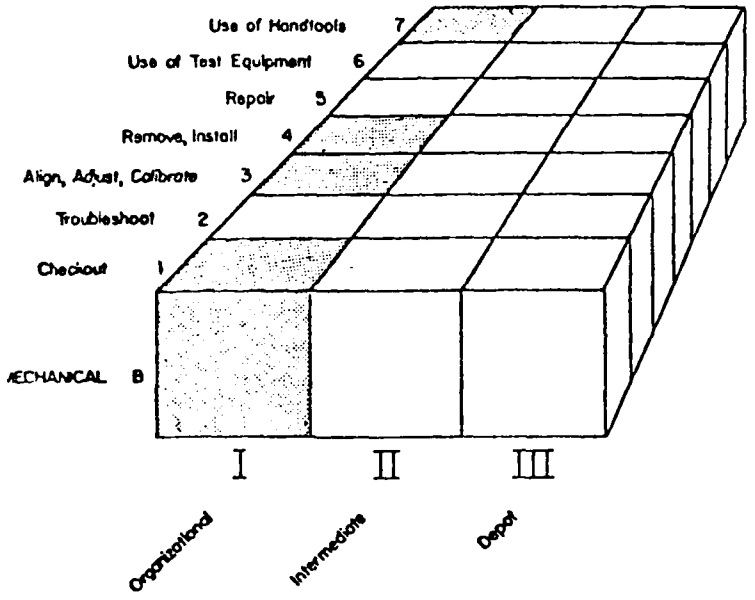


Figure 3 - A functional representation of the scope of the NumRAO PM development for mechanical maintenance (Vineberg et al, 1970b).

Found in Troubleshooting	Code	System Hardware Item	Reference Designator	Maintenance Function: Adjust	Align	Calibrate	Checkout/Troubleshoot	Clean	Disassemble/Assemble	Inspect	Lubricate	Operate	Remove/Install	Repair	Service	Notes
				1	2	3	4	5	6	7	8	9	10	11	12	13
✓	12	Control, Directional Listening C-8246	10	-	I*	-	I	I	I	-	-	-	O	I	-	*Resolver 1031 Alignment
	121	Knob		-	-	-	-	-	-	I	-	-	I	-	-	147971-1
	122	Panel, Control - Edge Lighted		-	-	-	-	-	-	I	-	-	I	-	-	159024-1
	123	Cover, Access		-	-	-	-	-	-	-	-	-	I	-	-	839691-801
	1231	Stud, Turnlock Fastener		-	-	-	-	-	-	I	-	-	D	-	-	2-0-100
✓	124	Amplifier, Driver-Directional Listening	10A2	-	-	-	D	-	-	-	-	-	I	-	-	718436-801
✓	1241	Capacitor	10A2 C15	-	-	-	-	-	-	-	-	-	D	-	-	CK06BX105K
✓	1242	Relay, Armature	10A2 K1	-	-	-	-	-	-	-	-	-	D	-	-	35BC1206A2
	1243	Insulator, Relay		-	-	-	-	-	-	-	-	-	D	-	-	7717-129N
✓	1244	Resistor	10A2 R1	-	-	-	-	-	-	-	-	-	D	-	-	RCR07G223J3
✓	1245	Resistor	10A2 R2	-	-	-	-	-	-	-	-	-	D	-	-	RCR07G223J3
✓	1246	Resistor	10A2 R3	-	-	-	-	-	-	-	-	-	D	-	-	RCR07G105J3
✓	1247	Semiconductor Device	10A2 CR5	-	-	-	-	-	-	-	-	-	D	-	-	JAN1N645
✓	1248	Resistor	10A2 R32	-	-	-	-	-	-	-	-	-	D	-	-	RNS5D1783F
✓	1249	Capacitor	10A2 C5	-	-	-	-	-	-	-	-	-	D	-	-	CM05FD221J03
✓	12410	Transistor	10A2 Q4	-	-	-	-	-	-	-	-	-	D	-	-	JAN2N930
✓	12411	Transistor	10A2 Q3	-	-	-	-	-	-	-	-	-	D	-	-	JAN2N930
	12412	Insulator, Transistor		-	-	-	-	-	-	-	-	-	D	-	-	10197DAP
	12413	Insulator, Transistor		-	-	-	-	-	-	-	-	-	D	-	-	10197DAP
✓	12414	Capacitor	10A2 C13	-	-	-	-	-	-	-	-	-	D	-	-	CK06BX104K

Figure 4 - Example of a Task Identification Matrix (TIM). Cell entries: - (dash) no maintenance task of this type is performed on this hardware item; O - task of type, performed at organizational level; I - task, performed at intermediate level; and D - task, performed at depot level.

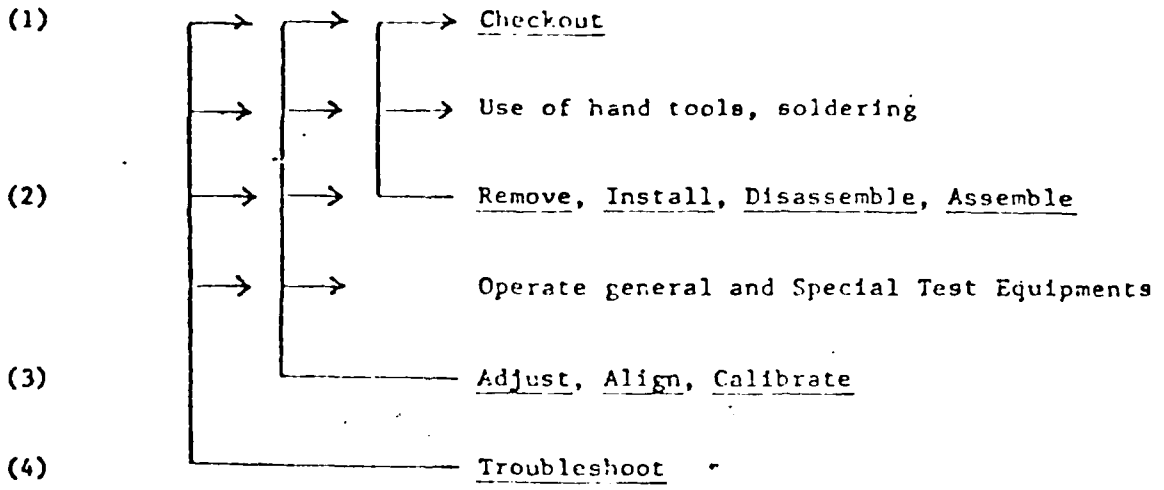


Figure 5 - Indicating the Dependencies among Maintenance Functions for an Electronic Hardware (Functions Underlined)

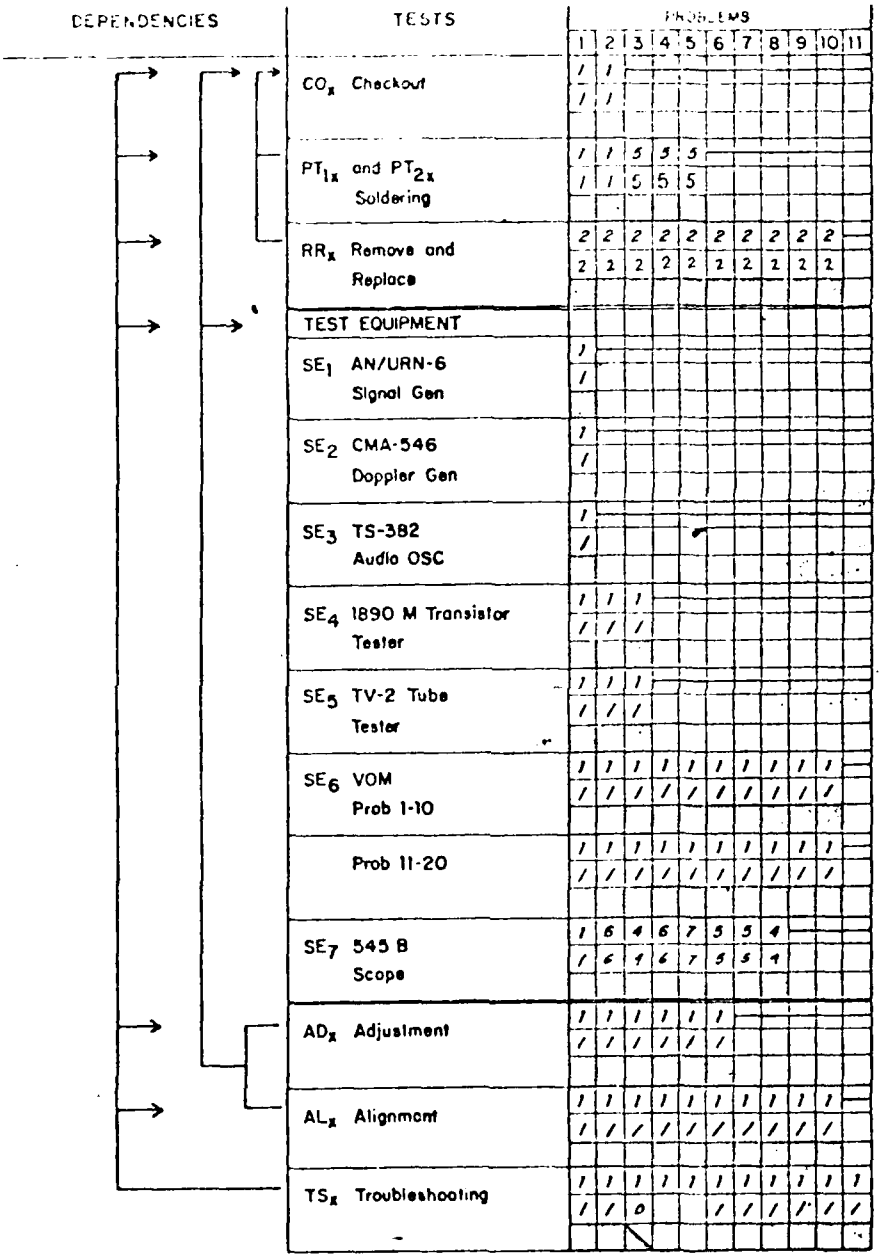


Figure 6. A profile for displaying the results obtained by an individual subject from a battery of Job Task Performance Tests concerning an Electronic System - the AN/APN 147 and the AN/ASN 35. This represents the profile of an individual who has successfully completed most of the battery.

REFERENCES

Anderson, A. V. Training, Utilization and Proficiency of Navy Electronics Technicians: III. Proficiency in the use of Test Equipment. Navy Technical Bulletin 62-14, AD-294 330. San Diego: U. S. Navy Personnel Research Activity, 1962.

Brown, G. H., Zaynor, W. C., Bernstein, A. H., and Shoemaker, H. A. Development and Evaluation of an Improved Field Radio Repair Course. Technical Report 58, Project Repair. AD-227 173. Washington, DC: Human Resources Research Office, The George Washington University, 1959.

Crowder, N., Morrison, E. J., and Demaree, R. G. Proficiency of O-24 Radar Mechanics: VI. Analysis of Intercorrelations of Measures. AFPTRC-TR-54-127, AD-62 115. Lackland AFB, TX: Air Force Personnel and Training Research Center, 1954.

Evans, R. N. and Smith, L. J. A Study of Performance Measures of Troubleshooting Ability on Electronic Equipment. Illinois: College of Education, University of Illinois, October 1953.

Foley, J. P., Jr. Critical Evaluation of Measurement Practices in Post-high School Vocational Electronic Technology Courses. AD-683 729. Doctoral dissertation, University of Cincinnati, 1967.

Foley, J. P., Jr. Description and Results of the Air Force Research and Development Program for the Improvement of Maintenance Efficiency. AFHRL-TR-72-72, AD-77 100. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, November 1973.

Foley, J. P., Jr. Evaluating Maintenance Performance: An Analysis. AFHRL-TR-74-57 (1), AD-A004 761. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, October 1974.

Foley, J. P., Jr. Factors to Consider in Developing New Test and Evaluation Techniques. Proceedings of the Human Factors Testing Conference, 1-2 October 1968. Snyder, M. T. (Chm.), Kincaid, J. P., and Potempa, K. W. (Ed.), AFHRL-TR-69-6, AD-866 485. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, October 1969.

Frederiksen, N. Proficiency tests for training evaluation. In R. Glaser (Ed.), Training Research and Education, Pittsburgh, PA: University of Pittsburgh Press, 1962.

Harris, D. and Mackie, R. R. Factors in Influencing the Use of Practical Performance Tests in the Navy. Navy Technical Report No. 703-1, AD-285 842. Washington, DC: Office of Naval Research, 1962.

Highland, R. W. A Guide for Use in Performance Testing in Air Force Technical Schools. ASPRL-TM-55-1, AD-65 480. Lowry AFB, CO: Armament Systems Personnel Research Laboratory, January 1955.

Jenkins, J. G. Validity for what? Journal of Consulting Psychology, March-April 1946.

Joyce, R. P., Chenzoff, A. P., Mulligan, J. F., and Mallory, W. J. Fully Proceduralized Job Performance Aids: Draft Military Specification for Organization and Intermediate Maintenance. AFHRL-TR-73-43 (I), AD-775 702. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, December 1973.

Mackie, R. R., Wilson, C. L., and Buckner, D. N. Practical Performance Test Batteries for Electricians Mates and Radiomen Developed in Conjunction with a Manual for Use in the Preparation and Administration of Practical Performance Tests. AD-98 239. Los Angeles, CA: Management and Marketing Research Corporation, June 1953.

Maier, M. H., Young, D. L., and Hirshfeld, S. F. Implementing the Skill Qualification Testing System. R&D Utilization Report 76-1, AD-A023 994. Arlington, VA: U. S. Army Research Institute for the Behavioral and Social Sciences, April 1976.

Saupe, J. L. An Analysis of Troubleshooting Behavior of Radio Mechanic Trainees. AFPTRC-TN-55-47, AD-99 361. Lackland AFB, TX: Air Force Personnel and Training Center, November 1955.

Scott, D. L. and Joyce, R. P. TEKTRONIX 545B Oscilloscope Training. AFHRL-TR-76-19, AD-A022 941. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (a)

Scott, D. L. and Joyce, R. P. TS-1100/U Transistor Tester Training. AFHRL-TR-76-20, AD-A022 930. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (b)

Scott, D. L. and Joyce, R. P. TS-148 Radar Test Set Training. AFHRL-TR-76-21, AD-A022 931. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (c)

Scott, D. L. and Joyce, R. P. TV-2A/U Tube Tester Training. AFHRL-TR-76-22, AD-A022 932. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (d)

Scott, D. L. and Joyce, R. P. URM-25D Signal Generator Training. AFHRL-TR-76-23, AD-A022 933. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (e)

Scott, D. L. and Joyce, R. P. 200 CD Wide Range Oscillator Training. AFHRL-TR-76-24, AD-A022 934. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (f)

Scott, D. L. and Joyce, R. P. 5245 L Electronic Counter Training. AFHRL-TR-76-25, AD-A022 939. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (g)

Scott, D. L. and Joyce, R. P. Fluke 803 Differential Voltmeter Training. AFHRL-TR-76-26, AD-A022 956. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (h)

Scott, D. L. and Joyce, R. P. HP-410B VTVM Training. AFHRL-TR-76-27, AD-A022 940, Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (i)

Scott, D. L. and Joyce, R. P. Kay Model 860 Sweep Generator Training. AFHRL-TR-76-28, AD-A022 957. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (j)

Scott, D. L. and Joyce, R. P. SG-299 B/U Signal Generator Training. AFHRL-TR-76-29, AD-A022 972. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (k)

Scott, D. L. and Joyce, R. P. Simpson 260 VOM Training. AFHRL-TR-76-30, AD-A022 984. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1975. (l)

Shriver, E. L. and Foley, J. P., Jr. Evaluating Maintenance Performance: The Development and Tryout of Criterion Referenced Job Task Performance Tests for Electronic Maintenance. AFHRL-TR-74-57 (II), Part I, AD-A004 845. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, September 1974. (a)

Shriver, E. L. and Foley, J. P., Jr. Evaluating Maintenance Performance: The Development of Graphic Symbolic Substitutes for Criterion Referenced Job Task Performance Tests for Electronic Maintenance. AFHRL-TR-74-57 (III), AD-A005 296. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, November 1974. (b)

Shriver, E. L. and Foley, J. P., Jr. Job Performance for UH-1H Helicopter: Controlled Field Tryout and Evaluation. AFHRL-TR-75-28 (I), AD-B006 295L. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, June 1975.

Shriver, E L., Hayes, J. F., and Hufhand, W. R. Evaluating Maintenance Performance: Test Administrator's Manual and Test Subject's Instructions for Criterion Referenced Job Task Performance Tests for Electronic Maintenance. AFHRL-TR-74-57 (II), Part II, AD-A005 785. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, January 1975.

Shriver, E. L., Hayes, J. F., and Hufhand, W. R. Evaluating Maintenance Performance: A Video Approach to Symbolic Testing of Electronics Maintenance Tasks. AFHRL-TR-74-57 (IV), AD-A005 297. Wright-Patterson AFB, OH: Advanced Systems Division, Air Force Human Resources Laboratory, July 1974.

Vineberg, R., Taylor, E. N., and Caylor, J. S. Performance in Five Army Jobs by Men at Different Aptitude (AFQT) Levels: 1. Purpose and Design of Study. Technical Report 70-18, AD-715 614. Presidio of Monterey, CA: Human Resources Research Organization, 1970.

Vineberg, R., Taylor, E. N., and Sticht, T. G. Performance in Five Army Jobs by Men at Different Aptitude (AFQT) Levels: 2. Development and Description of Instruments. Technical Report 70-20, AD-720 216. Presidio of Monterey, CA: Human Resources Research Organization, 1970.

Vineberg, R. and Taylor, E. N. Performance in Four Army Jobs at Different Aptitude (AFQT) Levels: 3. The Relationship of AFQT and Job Experience to Job Performance. Technical Report 72-22, AD-750 630. Presidio of Monterey, CA: Human Resources Research Organization, 1972. (a)

Vineberg, R. and Taylor, E. N. Performance in Four Army Jobs at Different Aptitude (AFQT) Levels: 4. Relationships Between Performance Criteria. Technical Report 72-23, AD-750 604. Presidio of Monterey, CA: Human Resources Research Organization, 1972. (b)

Wallace, S. R. Criteria for what? American Psychologist, June 1965. (a)

Wallace, S. R. The Relationship of Psychological Evaluation to Needs of the Department of Defense. Proceedings of 7th Annual Military Testing Association Conference. AD-681 096. San Antonio, TX: October 1965, 1-10.

Williams, W. L., Jr. and Whitmore, P. G., Jr. The Development and Use of a Performance Test as a Basis for Comparing Technicians With and Without Field Experience. The NIKE AJAX AFC Maintenance Technician. Technical Report 52, AD-212 663. Washington, DC: Human Resources Research Office, The George Washington University, January 1959.

ABOUT THE AUTHOR

Dr. John P. Foley, Jr. has been a research psychologist with the Advanced Systems Division (formerly, the Training Research Division) of the Air Force Human Resources Laboratory since 1962. He has been involved in exploratory and advanced development programs concerning Job Performance Aids, the Measurement of Job Performance, Job-Oriented Training, and Programmed Instruction. For 20 years prior to 1962, Dr. Foley was involved in Air Force Technical Training at Scott Air Force Base, Illinois and Lackland Air Force Base, Texas. Since his retirement from full-time employment with the Air Force Human Resources Laboratory in 1973, he has been working for the Laboratory on special projects on a part-time basis. A native of Ohio, Dr. Foley received an Ed.D degree as well as a B.Sc and B.Ed from the University of Cincinnati. He received and M.Ed from Our Lady of the Lake College at San Antonio, Texas. He is a member of the American Psychological Association, American Educational Research Association, Phi Delta Kappa, and Human Factors Society. Dr. Foley is author of more than 25 technical reports and articles.

SOME PROBLEMS IN TEAM PERFORMANCE MEASUREMENT

John J. Collins
Essex Corporation
Alexandria, Virginia

ABSTRACT

This paper discusses team performance and team performance measurement in terms of research capabilities, program support, operational support, and technical foundations and trends. Available evidence indicates little progress in developing this technology. Research and operational support on fundamental conceptual and technical problems about teams, team training, and team performance is not adequate and a pessimistic view is expressed regarding future progress. Suggestions are presented for capitalizing on advances in small group behavior research to facilitate the development of team performance technology.

This paper attempts to provide some perspectives on the state of the art or, perhaps more appropriately, the absence of the state of the art in team performance. The areas discussed include research capabilities, program support, operational support, and technical foundations and trends. The sources of information for these perspectives are literature-based and experienced-based. Both of these have influenced this writer's assessments, which are, to a large degree, ones of concern about progress to date and pessimistic about the rate of progress for the immediate future.

As a first point, I would assert that one of the principal problems characterizing team performance technology is in the area of professional expertise. More specifically, there is little evidence in the literature to indicate that even a small number of researchers are available who have been conducting team performance research over a period of time. Much like the job performance aids area, researchers come and go, performing usually on only one study and then shifting to some other technical area (Collins, 1977b). Support for this assertion is available from an examination of the references in Meister's (1976) comprehensive review (Chapter V) on team performance: There are 126 references and only 19 authors have more than one article. In addition, more than 80 percent of those studies were conducted during the 1950s and 1960s. These findings are very similar to those of recent research in team training, an area in which the Defense Science Board (1976) has expressed concern and recommended increased support. Other evidence, such as the limited number of disciplines that have been applied in conceptualizing and conducting both individual and team performance research, can also be cited. For example, most of the researchers are psychologists who have been trained and experienced in the psychology of individual behavior. One seldom finds a sociologist or social psychologist among them.

This is not to suggest that qualified individual research and support personnel are not available within government or in the academic and industrial communities. Rather, it is to say that the interdisciplinary teams necessary to investigate and solve team performance problems in all of their important dimensions have not been assembled and supported over time. As a result, we do not have an essential base of experienced researchers. I would suggest that these same conditions also exist--but perhaps to a lesser degree--in the area of individual performance.

Very much related to this first point is the second; that is, the priority assigned to team performance research and development in DoD Human Resources RDT&E. The relatively low priority assigned to team performance and the absence of a continuing program over the past several years undoubtedly accounts in large part for the limited number of team performance researchers. I would prefer not to speculate as to why this condition has existed and continues to exist. Whatever the reasons, the available data indicates little funding support for team performance research and development. Whether future program plans change this situation is a matter about which I have no information.

A third area involves problems in the conduct of team performance research in operational environments. Anyone who has tried to arrange for subjects, to obtain access to equipment, or to determine mutually-convenient schedules for conducting research in operational settings understands the difficulties involved. The Defense Science Board report also addresses the limitations of these opportunities as well as the availability of supporting analytical and statistical techniques. Even in the best of times, the problems have been difficult for both the operational commands and the research community. Now that operating conditions are becoming more difficult with increasing work loads and decreasing numbers of personnel, opportunities for team performance research in operational environments will probably not increase. I am not very optimistic about rapid progress in developing team performance technology unless more operational opportunities can be provided and some innovative approaches are developed for both laboratory and field research settings.

In addition to these administrative and operational problems are the scientific and technical influences and trends.

Glanzer and Glaser (1955) pointed out more than 20 years ago that little success had been achieved in developing methods for analyzing team performance. Obermayer et al. (1972) concluded that an economical means of objectively measuring team skills was still an elusive goal. The problem, of course, is not only team performance measurement. Rather, it involves the whole technology of team performance including such areas as team conceptualization, team objectives, team behaviors and events, and the conditions interacting with and affecting these behaviors, as well as a system of objective measurement and prediction. Obermayer et al. (1974) also point out as do other researchers (e.g., Larson, Sander, and Steinemann (1974) in a study of unit effectiveness measures) how purely subjective measures are in such areas as air combat maneuvering. Words like vague, too general, subjective are still being used to describe available measures and criteria. Larson et al. (1974) has suggested the use of the DELPHI technique as a timely systematic approach to extracting expert opinion. The technique undoubtedly has merit but is still highly subjective. Some progress has been made in developing observable event metrics, e.g., in the Tactical Advanced Combat Direction Electronic Warfare System (TACDEW), in casualty assessment techniques for Army Infantry exercises (Project REALTRAIN), and in SAGE (Project NORM), with its emphasis on situational factors as important influences of team performance (Wagner et al., 1976). Unfortunately, these are sporadic, uncoordinated contributions to a not-too-well defined matrix of needs and projects which has very few filled cells.

A very fundamental issue in all of this is the conceptualization of the team. Lorge, Fox, Davitz, and Brenner (1958) reviewed and categorized research relating to group and individual performance in group problem solving. These authors

invited attention to how groups were defined in terms of the research strategy and problems, and cautioned against generalizing from group to group without validating the underlying assumptions relating to each group. The "groups" discussed were the statisticized, climatized, concocted, ad hoc, and traditional. Lorge et al. viewed ad hoc groups as representing one end of the continuum and traditional groups as representing the other; that is, from the just-assembled to the well-established.

The statisticized group is concerned with aggregation rather than interaction. This group results from statistical computations (i.e., averaging of the products of independent and noninteracting individuals) and is not usually a group at all. Research on individuals making judgments that are then averaged to form groups is an example. Climatized groups are established on the basis of physical proximity. One form may involve interaction but no measure of group consensus; for example, in jury experiments. Another form may not provide for discussion but rather act as a sequel to individual evaluation; for example, a show of hands may be involved to show group choice. A third form may have neither interaction nor consensus; for example, in facilitation studies. The concocted group neither meets nor interacts. The unique elements of each individual's products are combined to form a so-called group product. Ad hoc groups are usually established for an experiment and cease to exist when they are completed. They are usually assembled to work together mutually and cooperatively on some externally assigned task and vary in the extent of cohesion and mutuality of purpose they achieve. The traditional groups are characterized by interaction of members over a period of time, and develop a "tradition" or working together for mutual and common purposes. This definition includes the concept of continuously emerging and becoming more cohesive, cooperative, committed, and productive (Lorge et al., 1958, pp. 337-340). Progress in viewing teams in the context of the changing dynamics of emergent (i.e., traditional) groups has been slow due largely to the emphasis on individual proficiency as the key to successful team performance, a position not substantiated in the literature.

More recently, Hall and Rizzo (1975) reemphasized that many difficulties stem from the fundamental problems of team definition and concepts. Without clear-cut concepts about what teams are, it is obvious that there is little information available about populations and sampling of teams. In the absence of such information, interpretation of whatever data is available on team performance in terms of representativeness or generalizability is impossible. I will discuss some suggestions later in this paper for dealing with these issues.

Team performance measures historically have attempted to emphasize productivity measures. However, such measures require consideration of the social aspects of organizations which influence productivity and very little attention has been given to these influences. Steiner (1972) has noted that there was a significant drop in interest and research in group productivity that only recently is beginning to change. Part of the explanation for this lack of interest is given by Hackman and Morris (1975) who note that, in spite of the thousands of studies of group performance, little is known about why some groups are more effective than others and even less about what to do to improve the performance of a given group working on a specific task. Much of the criticism of research on group performance centers on the failure to deal with the full input-process-performance sequence. Zander (1971) has also drawn attention to the need to deal with the questions of group motives and goals. His research findings, which in part have been confirmed by Bowen and Siegel (1973), point to the importance of how

individuals change from concern for self to commitments to the group and feeling responsible for its fate, arousal of members' desire for high degrees of group success, etc. Very little research has been conducted on these aspects of group performance leaving many unanswered questions. Davis (1969, 1973) also emphasizes that the development of group performance theory must consider group product, group structure, and group process, and that any attempt to segregate essentially overlapping and continuous phenomena rests on uncertain ground. Unfortunately suboptimum strategies for such investigations have been the most common focusing on one or two elements but not the three.

Another aspect of team and group performance that has received little attention has been the theoretical foundations related to team or group development. When 17-18-year-old men and women enter the Navy, attend schools to learn occupational skills, and are then assigned and reassigned over a period of 4-6 years as members of teams, it is clear a process of development, of both the individual and the teams of which he is a part, does in fact occur. There is evidence in the literature that team processes enhance proficiency so that the team becomes more than just the sum of individual proficiencies. We need more information about how the growth and development processes of teams operates; that is, we need research on a theory of team development.

The primary purpose of this paper is to discuss problems related to team performance rather than what is known about team performance. Meister (1976) presents 27 conclusions about team or group performance in his chapter on Team Functions (Chapter V). He points out that concepts about team processes and team training have been fuzzy and the results of research on determining whether training results in performance improvement have been disappointing. By way of a brief summary of problems, I would point to the following as illustrative of the many problems associated with team performance measurement and prediction.

1. Small number of experts in team performance.

2. Lack of continuing program support for this research.

3. Difficulties in conducting this research both in operational settings and in laboratories.

4. Absence of a theory of teams, team training, and team performance.

5. Absence of a growth and development orientation in the conceptualization of teams as units.

6. Narrowness of conceptualization of the dimensions of the operational and team environment.

7. Inherent weaknesses in measurement techniques (e.g., use of linear, additive psychometric models).

8. Failure to utilize what is known from related disciplines (e.g., organizational psychology, social psychology, sociology).

With reference to the last point, I recently completed a study of the potential contributions of small group behavior research to team training technology development sponsored by the Office of Naval Research (Collins, 1977a). A large

amount of information was collected and is presented in the report in terms of three levels of analysis: theory, methods and techniques, and findings on substantive variables. Let me cite a few examples of potential contributions described. The first relates to the concepts and definition of a team. Considerable research on the theory of group has been conducted for determining the necessary and sufficient conditions for the occurrence of group-like phenomena and for determining the defining variables for groups. A similar effort could be put forth to define operationally-useful concepts and parameters for various classes of teams, a taxonomy of teams developed, and population and sampling statistics obtained among other things. A second contribution comprises the contributions from research on group interaction, group motivation and development, and group productivity, which are providing analytical techniques, task typologies, and input-process-output models for studying teams as dynamic, growth-oriented units, the performance of which can be investigated in terms of organizational and system effectiveness. One important need which could be met from applications of these advances is an instructional system development model for teams. These and other advances are possible only if it is recognized that we must depart from the historical emphasis on the individual as the keystone for team performance and instead focus on the team in all its aspects.

REFERENCES

Bowen, D. D. and Siegel, J. P. Process and performance: A longitudinal study of the reactions of small task groups to periods perfotmance feedback. Human Relations, 1973, 26, 433-448.

Collins, J. J. A study of the potential contributions of small group behavior research to team training technology development. Alexandria, VA: Essex Corporation, 31 August 1977 (ONR Contract No. 00014-76-C-1076, NR 170-843)(a)

Collins, J. J. Some perspectives on the job performance aids technology base. Symposium Proceedings: Invitational Conference on Status of Job Performance Aids Technology, NPRDC TR 77-33. San Diego, CA: Navy Personnel Research and Development Center, May 1977. (b)

Davis, J. H. Group performance. Reading, MA: Addison-Wesley Publishing Company, 1969.

Davis, J. H. Group decision and social interaction: A theory of social decision schemes. Psychological Review, 1973, 80(2), 97-125.

Defense Science Board. Summary report of the task force on training technology. Washington, DC: Office of Director of Defense Research and Engineering, 22 February 1976.

Glanzer, M. and Glaser, R. A review of team training problems. ONR Technical Report, Pittsburgh, PA: American Institute for Research, 1955.

Hackman, J. R. and Morris, C. G. Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. Advances in Experimental Social Psychology, 1975, 8, 47-99.

Hall, E. R. and Rizzo, W. A. An assessment of U. S. Navy tactical team training. TAEG Report 18. Orlando, FL: Navy Training Analysis and Evaluation Group, March 1975.

Larson, O. A., Sander, S. I., and Steinemann, J. H. Survey of unit performance effectiveness measures. Technical Report 74-11, San Diego, CA: Navy Personnel Research and Development Center, January 1974.

Lorge, I., Fox, D., Davitz, J., and Brenner, M. A survey of studies contrasting the quality of group performance and individual performance 1920-1957. Psychological Bulletin, 1958, 55, 337-372.

Meister, D. Behavioral foundations of system development. Chapter V - Team functions. New York: John Wiley, 1976, 231-296.

Obermayer, R. W. et al. Combat-ready crew performance measurement system study. AFHRL Final Report of Contract No. F41609-71-C-0008. Williams Air Force Base, AZ: Human Resources Laboratory, May 1972.

Steiner, I. D. Group processes and productivity. New York: Academic Press, 1972.

Wagner, H., et al. Team training and evaluation strategies: A state-of-the-art review. SR-ED-76-11, Alexandria, VA: HumRRO, June 1976.

Zander, A. Motives and goals in groups. New York: Academic Press, 1971.

ABOUT THE AUTHOR

Dr. John J. Collins is a Vice President and Technical Director, Essex Corporation, Alexandria, Virginia. He has approximately 25 years experience in Behavioral Sciences research and research management. Prior to joining Essex Corporation in July 1974, Dr. Collins held various positions in Navy RDT&E and Operational Commands including the Navy Personnel R&D Laboratory, Washington, D. C., the Bureau of Naval Personnel; the Deputy Chief of Naval Operations (Manpower and Naval Reserve); the Office of Director, Research, Development, Test and Evaluation; and the Office of the Assistant Secretary of the Navy (Research and Development).

OPERATOR PERFORMANCE MEASUREMENT IN SYSTEM TEST AND EVALUATION

LCDR W. F. Moroney and LT W. R. Helm
Human Factors Engineering Branch
Pacific Missile Test Center
Point Mugu, California

ABSTRACT

System Test and Evaluation (T&E) is defined and three types of T&E are discussed. To the test and evaluation community, performance measurement implies system performance assessment, which is, in part, system effectiveness assessment. Performance measurement is therefore necessary in the determination of system effectiveness. Techniques and criteria for defining human limitations are described, as well as techniques to improve the integration of human capabilities into the total system performance spectrum.

Before the reader can appreciate the wide range of procedures, methodologies and techniques used to measure operator performance in Test and Evaluation (T&E), T&E needs to be defined. What is T&E? When does it begin? When does it end?

TEST AND EVALUATION

There are three types of T&E: developmental test and evaluation (DT&E), operational test and evaluation (OT&E), and production acceptance test and evaluation (PAT&E) (DoD Directive 5000.3, 1977). Authority for each type is delegated to a different organization.

Developmental Test and Evaluation

DT&E is test and evaluation conducted to (1) demonstrate that the engineering design and development process is complete, (2) demonstrate that the risks have been minimized, (3) demonstrate that the system will meet specifications for performance, compatibility, interoperability, reliability, maintainability, training, and logistics, (4) demonstrate that the system is suitable for service use, (5) provide test data and analysis that the developing agency (DA) needs to make modifications, and (6) certify that the system is ready for operational evaluation (OPEVAL). DT&E is planned, conducted, and monitored by the developing agency.

Operational Test and Evaluation

OT&E is test and evaluation conducted to estimate the prospective system's military utility, operational effectiveness, and operational suitability (including compatibility, interoperability, reliability, maintainability, and logistic and training requirements), and need for any modifications. OT&E will be conducted to determine the initial operational tactics which can most effectively utilize the demonstrated performance of the newly developed weapons system. OT&E will be accomplished by operational and support personnel of the type and qualifications of those expected to use and maintain the system when deployed, and will be conducted in as realistic an operational environment as possible. The Operational Test Force is the Navy-designated OT&E organization which is separate and distinct from the developing and procuring commands and from the using commands. OT&E is planned, conducted, and monitored by Commander, Operational Test and Evaluation Force (COMOPTEVFOR).

PRECEDING PAGE BLANK-NOT FILMED

Production Acceptance Test and Evaluation

PAT&E is test and evaluation of production items to demonstrate that the item procured fulfills the requirements and specifications of the procuring contract or agreements. It is the responsibility of each of the military departments and defense agencies (DoD component) to accomplish the necessary PAT&E throughout the production phase of the acquisition process.

Test and Evaluation Phases

Figure 1 delineates the complementary relationship between the three types of T&E throughout the life of a program.

Developmental Test and Evaluation (DT&E) is required for all acquisition programs and is conducted in four major phases:

1. DT-I is DT&E conducted during the conceptual phase to support the program initiation decision. It consists primarily of analysis and studies to derive the human factors/system requirements.

2. DT-II is DT&E conducted during the validation phase to support the full-scale development decision. It demonstrates that design risks have been identified and minimized. It consists of verifying the results of the special analysis and studies, including modeling and simulation on the critical areas identified earlier. It is normally conducted at the sub-system/component level, up to and including employment of engineering models for final evaluation.

3. DT-III is DT&E conducted during the full-scale development phase to support the first major production decision. It demonstrates that the design meets its specifications in performance, reliability, maintainability, supportability, survivability, system safety, and electromagnetic vulnerability. This phase may be further subdivided into additional phases, such as contractor technical evaluation (CTE) and formal Navy technical evaluation (NTE). The final subphase of DT-III is technical evaluation (TECHEVAL), the purpose of which is to certify that the design meets specified requirements and is ready for operational evaluation (OPEVAL).

4. DT-IV is DT&E conducted after the first major production decision to verify that product improvements or correction of design deficiencies discovered during OPEVAL, follow-on test and evaluation (FOT&E), or Fleet employment are effective.

Operational Test and Evaluation (OT&E) is required for all acquisition programs except for those programs designated by Chief of Naval Material. OT&E is subdivided into two major categories: initial OT&E (IOT&E, which is all OT&E accomplished prior to the first major production decision), and follow-on T&E (FOT&E, which is all OT&E after the first major production decision). OT&E is further divided into five major phases (3 IOT&E and 2 FOT&E).

1. OT-I is any IOT&E that may be conducted during the conceptual phase to support the program initiation decision. Most acquisition programs do not require OT-I. However, when an OT-I is conducted, existing systems or modifications thereto will normally be used to help estimate the military utility of the proposed system.

2. OT-II is IOT&E conducted during the validation phase to support the full-scale development decision. It provides an early estimate of projected operational effectiveness and operational suitability of the system, initiates tactics development, estimates program progress, and identifies operational issues for OT-III.

3. OT-III is IOT&E conducted during the full-scale development phase to support the first major production decision. OPEVAL is the final subphase of the OT-III. It consists of a demonstration of achievement of program objectives for operational effectiveness, operational suitability, and continuing tactics development. OPEVAL normally uses pilot production hardware and begins about 1 month after completion of TECHEVAL testing.

4. OT-IV is FOT&E conducted after the first major production decision but before production systems are available for testing. Normally, OT-IV is conducted with the same preproduction prototype or pilot production systems used in OPEVAL. OT-IV consists of the testing of fixes to be incorporated in production systems, completion of any deferred or incomplete IOT&E, and continuing tactics development.

5. OT-V is FOT&E conducted on production systems as soon as they are available. OT-V provides for a demonstration of the achievement of program objectives for production system operational effectiveness and operational suitability. In addition, OT-V includes OT&E of the system in new environments, in new applications, or against new threats.

Product Acceptance Test and Evaluation (PAT&E) is testing conducted on production items to demonstrate that systems meet contract requirements and specifications.

Human factors engineering personnel participate primarily in DT&E, to a considerably lesser extent in OT&E, and rarely in PAT&E.

Human Factors Engineering T&E Policy

Naval Material Command Instruction 3900.9 (1970) has established policies and requirements necessary to ensure adequate development of human factors aspects of systems and equipment under the cognizance of the Naval Material Command. The policy requires that the human element of Navy systems shall undergo the same development, test, and evaluation steps as equipment elements of the same system. This requires integration of appropriate human factors information into design and its use in all major management and/or technical decisions and documents.

Many human factors engineering specifications and standards provide guidance and criteria which are appropriate during certain phases of the acquisition cycle. Figure 2 illustrates how some of the many requirements relate to the various aspects of the acquisition cycle. A few of the most important specifications and standards from the T&E point of view are MIL-H-46855A (and the associated data items DI-H-2105 Human Engineering Test Plan and DI-H-2111 Human Engineering Test Report), MIL-D-8706B, MIL-D-8708B, MIL-D-23222A, MIL-M-8650B, MIL-M-18828A, and, of course, MIL-STD-1472B.

Thus, T&E is an integral part of the acquisition process and is not something which occurs after research and development (R&D). For a detailed review of what Human Factors Engineering (HFE) does during system T&E, the reader is referred to Holshouser's "Guide to Human Factors Engineering General Purpose Test Planning (GPTP)" (1977). Now that the reader has a general feel for the "Why, What, and When" of T&E, let us attempt to describe performance measurement.

PERFORMANCE MEASUREMENT

To the test and evaluation community, performance measurement implies system performance assessment, which is, in part, system effectiveness assessment. Therefore, to the T&E community, performance measurement is a necessary part of the process of determining system effectiveness. System effectiveness can be viewed as a measure of the extent to which a system can be expected to complete its assigned mission within an established time frame under stated environmental conditions; it includes reliability, maintainability, logistical and technical support, and subsystem performance of man and machine, as well as problems associated with threat and mission analysis.

At Pacific Missile Test Center (PACMISTESTCEN) the human factors specialists conduct test and evaluation on weapon systems to determine a measure of effectiveness. To many people the words "weapon system" brings to mind a collection of sophisticated electronic components (such as make up a radar or an infrared system) used to discover the presence of unfriendly targets, and the weapon itself, which will be guided toward and used to destroy a target. It consists of many subsystems: an electrical subsystem, a target detection subsystem, a target classification system, a weapon platform system, a weapon launch and propulsion subsystem, and a target destruction subsystem. In an "air-launched" weapon system, the platform itself (e.g., an aircraft) is airborne at the time of launch. The concept of weapon system, therefore, must be enlarged to include all those subsystems necessary to launch the platform itself (such as a catapult system), to keep the platform in the air (a power plant system), to maneuver the platform to and from the general target area (a navigation, avionics, and flight control system), and to recover the platform when the mission is complete (a recovery system).

So far, only some of the "hardware" systems which are necessary for an air-launched weapon system have been mentioned. The total weapon system also includes a trained pilot and crew who are responsible for monitoring the status of the various in-flight subsystems previously mentioned and who constitute the "decision" subsystem in both the "target engagement" phase and the "platform maneuvering and control" phase of the total mission. With few exceptions, the functions which are performed by an operator in a weapon system are decision functions. It is generally agreed that operators spend a good portion of their time observing displays and manipulating controls, but the reason they observe displays is to get the information necessary to make a decision, and the reason they manipulate controls is to inform the rest of the system of their decision.

It is the operator and his decision-making function toward which a majority of our human factors investigations at the PACMISTESTCEN are directed.

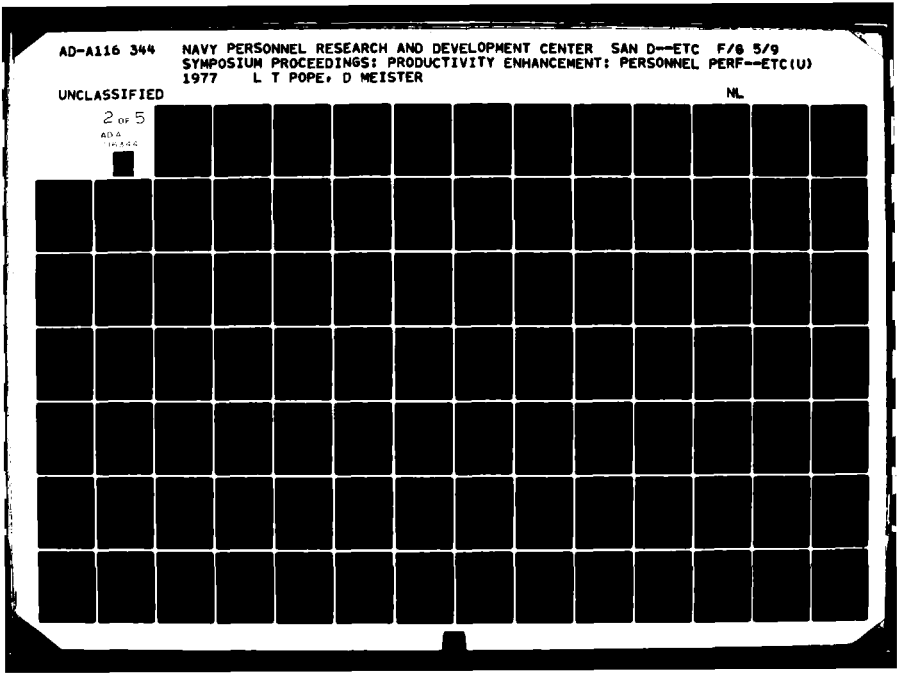
TEST AND EVALUATION CRITERIA AND METHODOLOGY

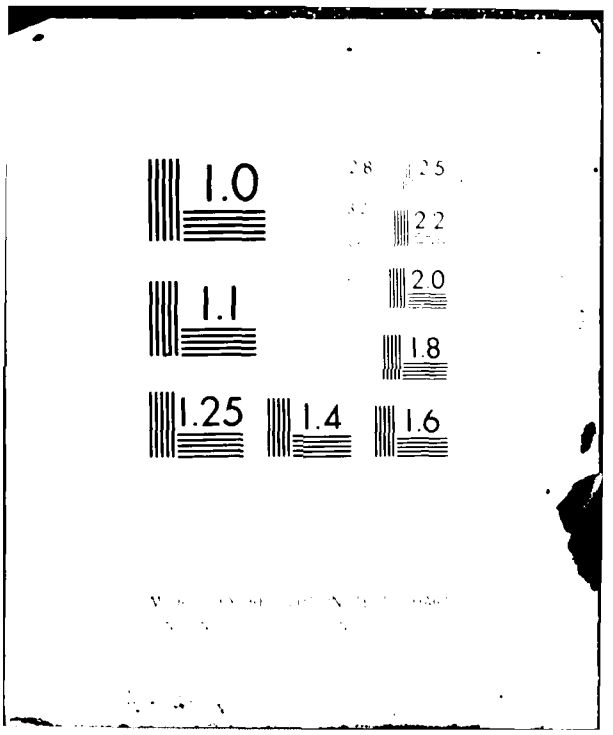
Although our assessments focus on the operator/decision maker subsystem, reasonable cost effectiveness evaluations and design trade-off considerations are dependent upon the ability to estimate total system performance capabilities, which include the impact of man-machine interactions. Inadequate human factors information during these trade-off evaluations leads to serious overestimation of fleet operational capabilities because these systems contain technical capabilities which cannot be used effectively by the available operators. Safety and combat effectiveness are seriously degraded by the presence of design deficiencies that increase the probability of inefficient use and improper maintenance.

AD-A116 344 NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 5/9
SYMPOSIUM PROCEEDINGS: PRODUCTIVITY ENHANCEMENT: PERSONNEL PERF--ETC(U)
1977 L T POPE, D MEISTER

UNCLASSIFIED NL

2 OF 5
AD A
116344




1.0
1.1
1.25
1.4
1.6
2.2
2.0
1.8

The development, implementation, and evaluation of human factors requirements during system development test and evaluation is made mandatory by DoD and Navy policies and instructions; however, currently available methodologies and criteria are insufficient for assuring their complete accomplishment. The implementation of human factors considerations during systems development test and evaluation is largely dependent upon discrete translations of standards, specifications, and conventions into engineering design configurations, with little provision for testing their validity or sufficiency in any particular application. Information concerning the bases for human engineering decisions and their anticipated impact upon performance is usually not available during later development phases, trade-off analyses, or test and evaluation. During operational test and evaluation, the constraints of time and resources preclude the reconstruction of human factors assumptions and considerations made during development, and reduce the human factors test and evaluation (T&E) effort essentially to counting knobs and dials and sampling operator opinions. In spite of these difficulties, there is a substantial technical base that can be used for developing the methodologies for the identification of critical test points, as well as for defining the necessary performance criteria required during systems development.

To overcome the lack of insufficient design criteria and evaluation methodology, the Human Factors Branch at the Pacific Missile Test Center (PACMISTESTCEN), under the sponsorship of Naval Air Systems Command (NAVAIRSYSCOM), has initiated and continued two separate but interrelated efforts. One effort concentrates on the verification and assessment of design criteria while the other centers on the improvement of test and evaluation methodology.

The first effort seeks to avoid exceeding man's limitations by providing test techniques and adequate design criteria to be used during the system development, test, and evaluation cycle. The second effort addresses the integration of man's capabilities into the total system performance spectrum by providing the tools, techniques, procedures, and methodologies necessary in total system evaluations. In combination, these efforts seek to estimate system/subsystem technical performance and its relation to the system's operational worth. Each of these efforts and their associated techniques will be discussed in subsequent sections.

Design Criteria: Verification/Assessment and Related Techniques

As part of the design criteria effort, PACMISTESTCEN (1) has participated in the revision of a number of standards/specifications, and (2) has collected and published human performance data applicable to design and T&E. Closely related to the above is the development of techniques to apply in assessing contractual compliance and impact on system effectiveness. The techniques used in assessing performance during T&E vary as a function of where the weapon system is in the acquisition cycle, because different levels of information and equipment "hardness" exist at different points in that cycle. A recent review by Geer (1977) described T&E techniques. A detailed description of these techniques would be inappropriate in this paper, so abstracts of selected techniques are presented below:

Human Factors Test and Evaluation Manual (HFTEMAN)(Malone and Sheak, 1976)--HFTEMAN is designed to assist the HF engineer in the areas of test plan preparation, test conduct, test data evaluation and analysis, and test report preparation. the HFTEMAN consists of three documents: the first contains detailed HFE test data, the second is a supplement that contains specific HFE design criteria, and the third describes methods and procedures usable in HF T&E.

The procedure of using HFTEMAN may be considered as a five-step process: the first step requires that test item be classified as a vehicle, weapon, electronic equipment, etc. The second step is to identify both the user functions and the tasks related to this type of equipment; in other words, a selection is made of what to evaluate and the criteria to be used in the evaluation. The third step decides what human factor considerations and what item components are relevant. The test observer then reviews the task list to identify which of the test item components apply to the item under test, and which human factors considerations are important. In the fourth step, on the basis of the identified equipment components and the identified human factor considerations, the HF engineer enters the appropriate row and column of a matrix. The cells of this matrix contain the exact test criteria. In the last step, these criteria are used to prepare the HFE test plan.

HFTEMAN may be used, in various levels of detail, on any program at any time during the program evolution. It provides both the basis on which to build a HF checklist and the necessary information for HFE T&E planning and conduct.

Computer Accommodated Percentage Evaluation (CAPE)--Workspaces traditionally have been designed without knowledge of the proportion of the user population that is accommodated with safety and full capability. In aircraft cockpit design, for example, designers have been directed to develop cockpits that accommodated 5th through 95th or 3rd through 98th percentile operators. However, crew systems designers usually consider only one anthropometric feature at a time and ignore the interaction between variables. The combination of all the necessary dimensions that make up a workspace design limit the operators to a much smaller actual range than expected. It has been shown (Moroney and Smith, 1972) that more than 50 percent of the 1964 population of naval aviators would be excluded when 5th and 95th percentile critical limits are imposed for 13 cockpit-related variables. When the 3rd and 98th percentile values are used, over 32 percent were excluded. This problem has led to the development of CAPE. The CAPE is a Monte Carlo computer model for generating representative pilot anthropometric features (including links) and comparing these data with an adjustable work-space model so that the population accommodated by the workspace can be estimated and maximized. The CAPE model has two options: exclusion demonstration and workplace analysis. Each option, and its underlying model with components, is described in summary form below. More detailed descriptions of model options, their components, and the total CAPE model are contained in Bittner's report (1975).

In CAPE, the exclusion demonstration option determines what percentage of the potential population is excluded from a workspace design with respect to each anthropometric feature entered into the program. This option may be considered to be composed of two components--an exclusion limits component and a Monte Carlo sample generator--but only the former will be discussed here.

The exclusion limits component provides for the entry, storage, and utilization of user-provided standard score limits of anthropometric variables required for exclusion studies. For each variable involved in an inclusion demonstration analysis, high cutoff and low cutoff values must be input by the user. This component of the analysis provides for the sequential testing, element by element, of Monte Carlo-generated standard score vectors to determine if the vectors are within the limits set by the high and low standard score boundaries (populations

of standard scores have means of zero and standard deviations of one). Rejection of any component test is defined as nonaccommodation of that (sample subject) feature vector.

The workplace analysis option determines the percentage of a population that will be excluded from a cockpit design based on the geometric parameters of the workplace. The workplace analysis option of the CAPE program can be thought of as being composed of four components: (1) an operator link system component, (2) a sample operator generator component, (3) a component characterizing a seat-workspace layout, and (4) a workspace testing component. The operator link system is an abstraction of a pilot's anthropometry. The sample operator generator provides sets of operator link values suitable for input in the operator link model. When the compatibility of a geometry is desired, these values are entered as the mean, and the generator provides only this one sample. However, when an operator accommodation analysis is desired, this generator provides quasirandom samples from a multivariate normal population.

Selected HFE T&E Reports--As part of the design criteria effort, a number of reports relating human performance data to design criteria and T&E have been prepared. Some of these reports are discussed beiefly below:

Assignment of females to a wider variety of previously all-male occupations led to the discovery that, while a particular cockpit accommodated 88 percent of the male population, only 11 percent of the female population were accommodated (Ketcham-Wiedl and Bittner, 1977). This finding is based on the CAPE model described previously.

Problems encountered in the area of workplace design led to the preparation of a report (Ayoub and Halcomb, 1976) which contained an annotated bibliography and utilized CAPE to determine the percentage of the population excluded from a workplace.

The debilitating effects of motion sickness encountered in the air or sea environment seriously degrade an operator's ability to perform his mission, yet no design limitations are contained in military specifications which define acceptable regions for the human operator or inform the designer of such vehicles of the expected incidence of motion sickness. Therefore, using motion sickness incidence data collected under Office of Naval Research funding, an effort was undertaken which resulted in reformatting the data into a model for specifying operator limitations for design considerations (see McCauley and Kennedy, 1976).

Currently, the effect of heat stress on operator performance is being reviewed and will result in a procedure/document which will allow us to approximate the expected change in performance as a function of temperature. Another ongoing effort is designed to define for the R&D community which anthropometric features need to be measured in a new survey of our naval aviator population.

Test and Evaluation Methodology

Whereas the PACMISTESTCEN effort in verification and assessment of design criteria concentrates on the identification of system parameters where there are certain limitations to the performance capabilities of the human operator which no amount of practice or training will overcome, the test and evaluation methodology improvement effort seeks to estimate the range of man's capabilities and his overall contributions to system effectiveness within a total system perform-

ance framework. This effort deals with the concept of task loading, that is, the measurement and assessment of an operator's contribution to system effectiveness in terms of his capabilities and specific system parameters. Although systems specifications require the analysis of task loading imposed on the operator, the definition of "task loading" is little understood.

It has been discovered in testing and evaluating many current aviation systems that the operators are required to perform complex tasks under excessive pressure of environment- and task-induced stress. It has been determined that the effectiveness of these systems is dependent upon the operator's capacity to process and respond to a large quantity of information. The methodological problem is that there are no adequate measurement techniques for quantifying either human workload capacities or system demands made upon these capacities. Faulty techniques and misinterpretation of available data can lead to the development and deployment of systems in which the operator is severely overloaded and is required to perform near-impossible sequences of perceptual, cognitive, and manual tasks.

The test and evaluation methodology effort is developing techniques to quantify operator workload capacities and system operability indices to provide methods for the quantitative assessment of the effects of tasks and of environmental and operator variables upon total system effectiveness. Each of the techniques undergoing development will be described below, but, before proceeding, it would be beneficial to explain our concept of evaluation and how we think it relates to operator capabilities, system performance, and operational military worth.

First, any technique developed for assessing workload or system operability is, above all, an evaluating mechanism. The logic of evaluation requires that any technique for evaluating system/subsystem design and performance parameters must be carefully constructed to meet four general criteria:

1. It must discriminate effectively among the alternative design parameters.

2. It must be reliable.

3. It must be intelligible (i.e., have an explicit logic that facilitates understanding of the relationships between the data and the results of the evaluation and be equitable).

4. It must be equitable (i.e., have no inherent bias).

Any technique for evaluation, however constituted but meeting these criteria, has but one operational purpose: to relate available data relevant to the evaluation of a system to its total operational worth to the Navy. The question of the system's operational worth actually raises two distinct but related evalution questions: what is the actual performance of a system, and what is that level of actual performance worth? The first question requires some quantifiable assessment of system performance that can be used to predict actual performance. The second question deals more with evaluation in its broadest sense. Answers to questions requiring a prediction concern the effectiveness of a system; answers to questions requiring an evaluation concern the utility of that system. While these distinctions between prediction and evaluation and between effectiveness and utility are observed in the development of our evaluation methodology techniques, the problem to which such methodologies address themselves will be regarded as one of evaluation in a broader sense.

There are three methodologies that are being developed to answer the questions of system performance and operational military worth: the first is the development of operator workload assessment techniques that can be used to predict operator task loading. The second technique, the Function Description Inventory (FDI), is used to estimate levels of task difficulty and subsystem effectiveness within the hierarchical task structure of a specific mission. The third technique combines elements of multiattribute theory and the FDI approach and is used to estimate task operability across a mission profile. Each of these techniques will be discussed in more detail below.

Workload Assessment Techniques--PACMISTESTCEN, in conjunction with the Naval Air Test Center and Wright-Patterson Air Force Base, is pursuing the development of objective workload assessment techniques. In the past, workload was defined in the context of crew work-rest schedules or time-and-motion studies. It was typically measured by "amount of expended physical effort" or "time analysis of activities." However, these types of approaches were found to have the same methodological problems as some of the more modern techniques of measuring workload, for example, man-simulation models, because they are not applicable in the test and evaluation environment. Regardless of how workload is to be measured, it must ultimately be directly relatable to the evaluation of aircrew systems in an operational setting if it is to have utility and validity for the test and evaluation community. Therefore, an effort was initiated to compile an annotative bibliography of methodologies that measure operator workload in aircrew systems (Schiflett, 1976). The result of this review effort was to compile and catalog effective and easily adaptable analytical methods for use in assessing operator workload in the test and evaluation process. It was assumed that any source of information or any method developed for operator workload assessment and/or prediction would have some utility for the test and evaluation community. It was found, however, that the majority of methods used for workload assessment were developed as an aid in the design of aircrew systems; consequently, the methods are difficult and/or impractical to implement in the test and evaluation environment.

Based on the results of this review, it was determined that future development of workload measurement methodologies need to be systematically evaluated and integrated into the context of real-world, complex aircrew systems. A taxonomy matrix of basic operator activities was recommended to classify generic aircrew tasks. Currently a contract has been let to conduct a comprehensive state-of-the-art survey and analysis of workload measurement methodologies to identify techniques specifically applicable by the test and evaluation community. The results of this survey will be available in the summer of 1978. Based on the survey's recommendations, workload measurement techniques will be flight-tested at the Naval Air Test Center with the ultimate goal of obtaining objective operator workload capacity indices to be used in predicting inflight performance.

Function Description Inventory--The Function Description Inventory (Helm, 1975) is a tool for providing operator-based, quantifiable assessments of the effectiveness of man-machine compatibility and is an aid toward integrated subsystem analysis in the total weapon system context. The procedure for FDI development and employment requires a series of investigations analyzing the operational functions of an operator, with an essential part involving the determination of the roles, duties, and tasks (these will be explained in detail below) performed by a crewmember. Next, averages of crewmember judgments are compiled on how important these roles, duties, and tasks are for mission success; how frequently they are performed on a typical mission; how difficult it is to perform the task; and, finally, how effective the weapon system is in accomplishing the operational

functions. Analysis of roles, duties, and tasks across these four dimensions provides a considerable degree of in-depth evaluation of the interrelated problems within the man-machine system and gives an additional perspective that is usually not available through discrete analysis of human engineering design deficiencies.

The approach combines some of the best features of the checklist, open-ended questionnaire, and interview methods. The principal tool used in this procedure is an inventory of activities. The inventory method has the advantage of being a simple procedure in that the crewmembers merely check their rating of the roles, duties, and tasks listed and, if necessary, write in those duties and tasks which do not appear. The procedure is economical in that it allows for a broad sampling, a ready synthesis of a large amount of information, and a high level of standardization, and lends itself to computer analysis of the data provided by the inventory.

In function analysis, operational mission activities are categorized into the following three hierarchical levels:

1. Role--A broad category of activity performed by a crewmember. Each role may encompass a number of duties and tasks. These roles encompassed 100 percent of the responsbilities of the crewmember within an operational mission framework.

2. Duty--A large segment of activity performed by a crewmember. All duties under a role in combination define 100 percent of the role.

3. Task--A unit of work activity which forms a significant part of a duty. All the tasks under a duty in combination define 100 percent of the duty.

The developed FDI is then administered to experienced fleet operators for their ratings of task criticality, frequency, difficulty, and system effectiveness. Data are then analyzed by computer. Computer analysis involves generating and/or computing frequency distributions, means, standard deviations, percentages, and rank ordering. From the computer analysis, summary tables presenting the rank order of roles by mean value of criticality of activity, frequency of performance, difficulty of task and system effectiveness can be presented in tabular form. These values are derived from each crewmember's ratings of each role, duty, and task on each dimension. For a detailed examination of this procedure, see Helm (1976a).

The initial use of the FDI as a T&E tool was accomplished as part of the S-3A test trials (Helm, 1975). Since then, the FDI has been used to evaluate operator positions in the P-3C and E-2C aircraft. In addition to using the FDI methodology on fleet aircraft, a validation effort was undertaken by comparing recorded engineering design deficiencies in the S-3A aircraft against FDI indicators of potential human factors problems (Helm, 1976b). There was a high agreement between noted human engineering deficiencies and low ratings in system effectiveness by FDI crewmembers. This result gives considerable support to the belief that the FDI is a valuable human factors tool in assessing man-machine compatibility in complex weapon systems.

Mission Hierarchy Analysis--In the test and evaluation phase of system acquisition, large amounts of data are available. The different pieces of information have different relative importance in terms of the implications for required decisions. These data must be organized in a manner that accomplishes certain functions. One function is to monitor various aspects of system performance to

facilitate decisions, for example, how good the system is with respect to such issues as: the quality of life support systems, personnel fatigue factors, subsystem operability, and subsystem effectiveness.

These issues are not addressed by single pieces of the available data, but, somehow, the data must be organized to facilitate meaningful statements about such factors. A mechanism is necessary to combine very precise pieces of test information into summary measures at differing levels of generality to facilitate decisions about possible alternatives.

The methodology being developed at PACMISTESTCEN to satisfy these criteria involves the integration of two existing techniques: the first has been used extensively in human engineering task or job analysis. Task analysis provides the information about operator activities relevant to the operation of a complex system. The second technique is more recent and involves the implementation of a hierarchical multiattribute utility model.

Combining these two techniques results in a methodology that takes as inputs operator activities and related pieces of specific data and organizes this information into a hierarchical structure so that each successive higher level in the hierarchy is of a greater degree of generality. With this methodology, operator assessments of task difficulty and subsystem effectiveness can be integrated and combined into a structure consistent with the rules employed in multiattribute utility theory such as combination rules, importance weighting, and utility functions. Utilization of this approach should provide those involved in human engineering trade-off analysis with data about the difficulty of specific operator activities and estimates on subsystem effectiveness within a hierarchical structure that can relate data from the general to the specific level within the system.

Preliminary investigations using the operator control functions in the F-18 indicate that it is possible to combine task analytic techniques and a multiattribute utility model to achieve a comprehensive assessment of a complex system (O'Connor and Buede, 1977). To test this approach, the complete F-18 system currently undergoing development has been selected as a test vehicle. Using a mission/task hierarchical approach as illustrated in Figure 3, a preliminary task analysis and an evaluative computer program for the F-18 system have been developed. The pyramidal structure is illustrative of this approach. The tasks required to operate the F-18 have been arranged in a mission phase structure so that estimates of operability can be obtained at the lowest mission level. These estimates are then successively integrated throughout the hierarchy. The F-18 task analysis and computer evaluative program will be available by Summer 1978. Using this approach, data will be collected during F-18 flight testing and the results will be presented to the F-18 program manager through an interactive computer system. Thus, the program manager can determine overall system evaluations and, more important, how specific operator activities have contributed to total system operability. With this information, the program manager can determine what changes to the F-18 will be most beneficial in improving system operation.

NEW DIRECTIONS

Both the verification and assessment of design criteria and the HFE T&E Technology efforts described above have and will lead to improved Human Factors T&E techniques. The task of system performance measurement, however, doesn't stop when performance has been measured. For the T&E community, the best performance assessment techniques are useless unless they have impact on the decision-making process relevant to systems undergoing acquisition. There are basically two ways to impact on decision making: one is legal and official, the other is unofficial. The former is harder but sometimes necessary, while the latter is easier and less costly in terms of time, money, and effort. The effectiveness of the unofficial approach is a function of the program/project manager's (PM) familiarity and previous experience with human factors. If he has had positive exposure to and/or formal education in human factors (at least an introductory course), the probability of impacting his decision making and, ultimately, the system increases significantly. Unfortunately, without pervious exposure to human factors, PMs often view HF as an add-on, nice to have if there is enough money available. To counter this, as part of our T&E Technology work unit we have undertaken a communication effort to develop operationally oriented films and manuals that convey to operational and managerial personnel the importance of good human factors design, test and evaluation. It is through this effort that we hope to "educate" those involved in decision making and in the operation of new and modified systems on the availability and advantage of human factor tools, techniques, and procedures for the assessment of system performance and ultimate system operational effectiveness.

SUMMARY

In this paper we have described the nature of T&E, equated performance measurement with the determination of system effectiveness, described selected techniques and criteria that have been developed to better define human limitations, described techniques which have been developed to improve the integration of man's capabilities into the total system performance spectrum, and, finally, addressed the need to communicate the contribution of human factors to management personnel. We hope that we have expanded the reader's perception of performance measurement during the test and evaluation of naval systems.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the contribution of Mr. Edward Holshouser of the Pacific Missile Test Center for the use of material describing the Test and Evaluation process. Several concepts developed by Commander R. Wherry, USN (Retired), former head of the Human Factors Branch at the Naval Missile Center, have also been woven into this paper, and his contributions are also acknowledged.

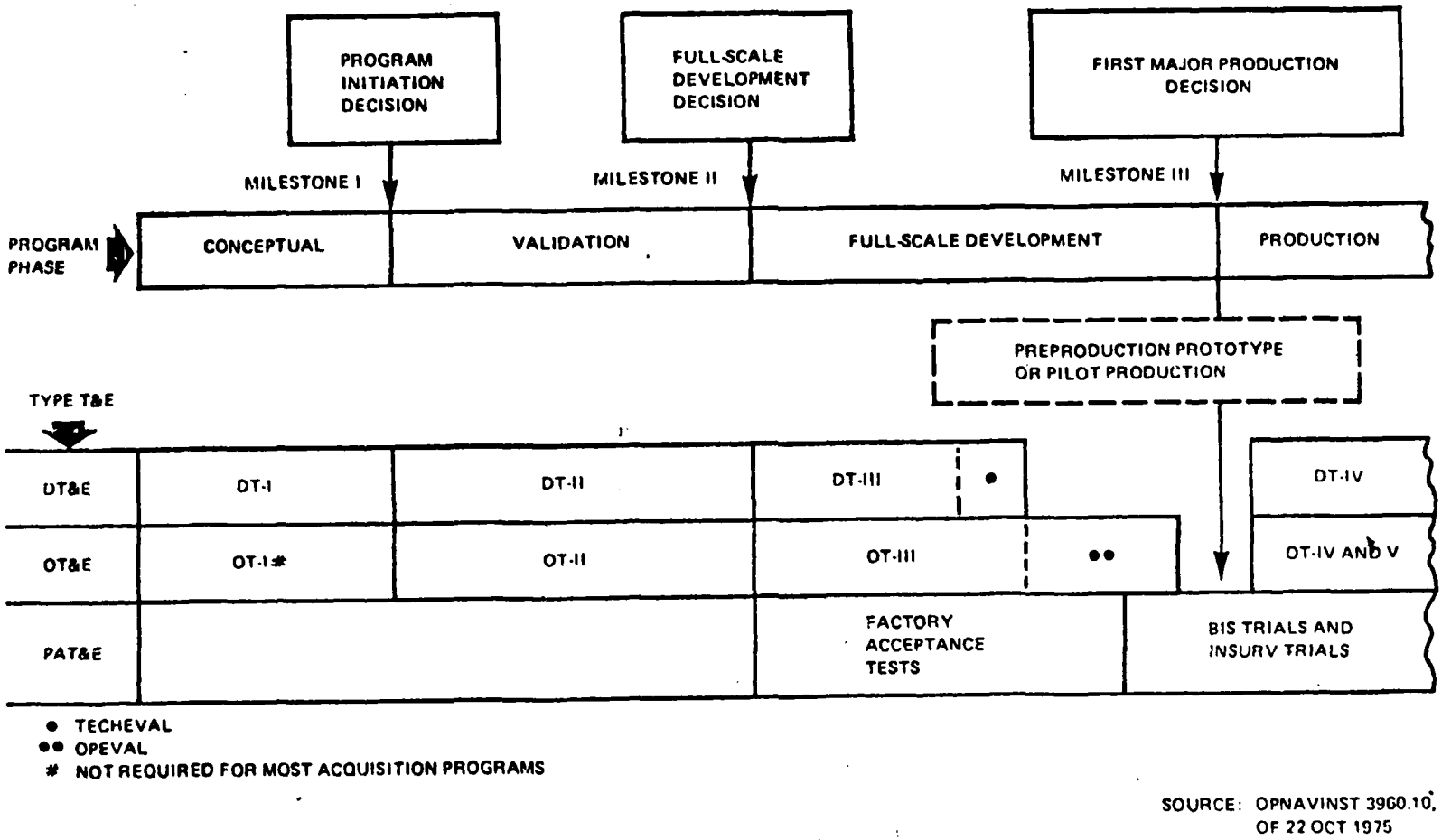


Figure 1. Test and Evaluation phases.

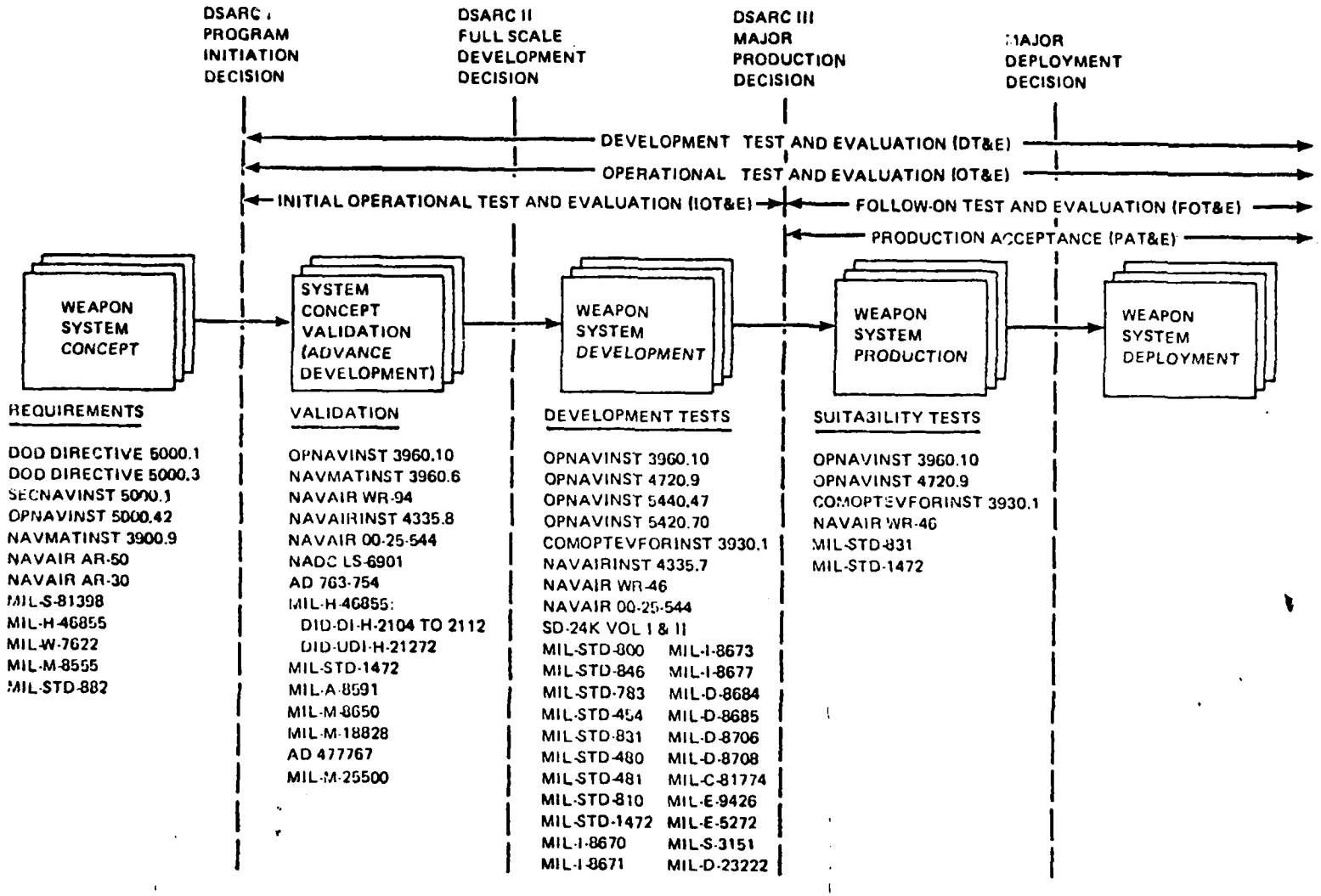


Figure 2. HFE System Requirements.

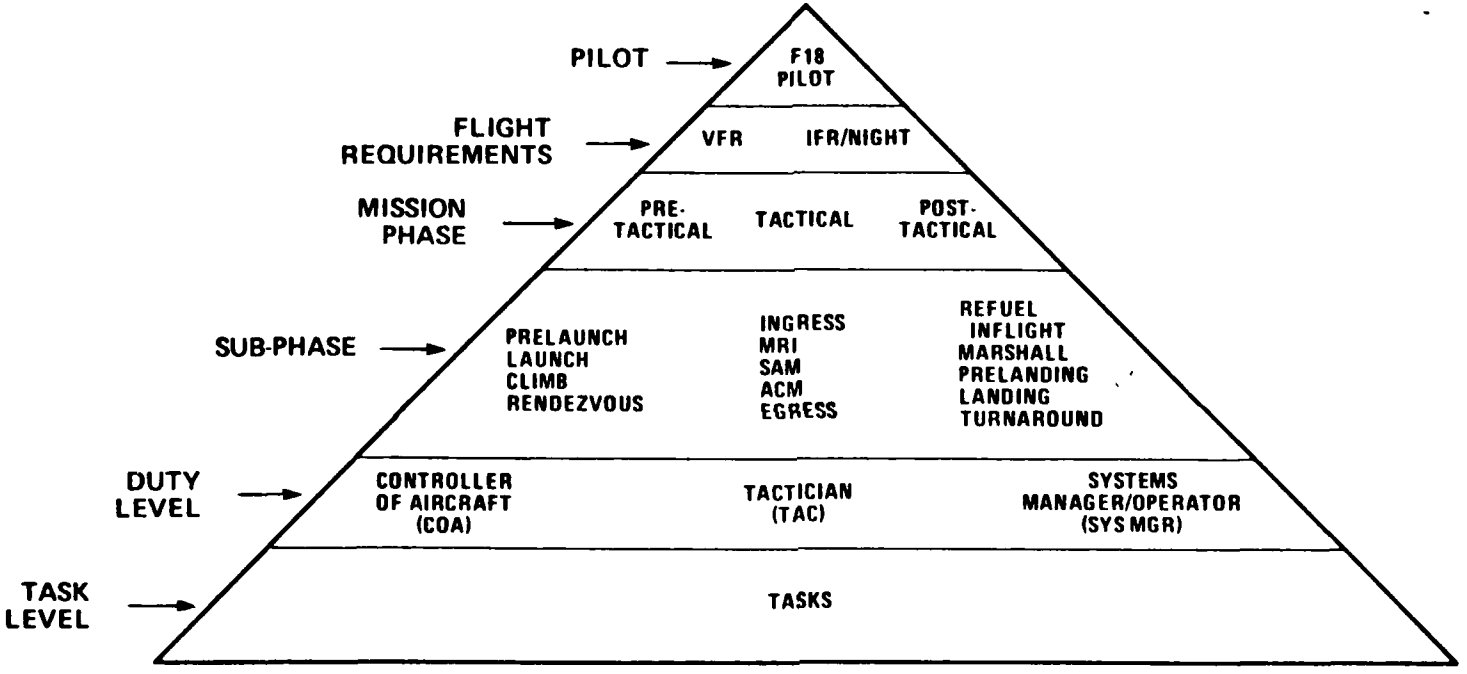


Figure 3

THE F18 PILOT TASK INVENTORY HIERARCHY

REFERENCES

Ayoub, M. M. and Halcomb, C. G. Improved seat, console, and workplace design: Annotated bibliography, integration of the literature, accommodation model, and seated operator reach profiles. TP-76-1, Point Mugu, CA: Pacific Missile Test Center, December 1976.

Bittner, A. C., Jr. Computerized accommodated percentage evaluation (CAPE) model for cockpit and other exclusion studies. TP-75-49/TIP-03, Point Mugu, CA: Pacific Missile Test Center, December 1975.

DoD Directive 5000.3 "Test and Evaluation." Department of Defense, Washington, DC: March 7, 1977.

DoD, MIL-H-46855A, Human engineering requirements for military systems, equipment, and facilities. Washington, DC: May 2, 1972.

DoD, Data Item Description, DI-H-2105, Plan, human engineering test. Washington, DC: July 20, 1973.

DoD, Data Item Description, DI-H-2111, Report, human engineering test. Washington, DC: July 20, 1973.

DoD, MIL-D-8706B, Data and tests, engineering: Contract requirements for aircraft weapon systems. Washington, DC: August 15, 1968.

DoD, MIL-D-8708B, Demonstration requirements for airplanes. Washington, DC: January 31, 1969.

DoD, MIL-D-23222A, Demonstration requirements for rotary wing aircraft (helicopters). Washington, DC: March 18, 1971.

DoD, MIL-M-8650B, Mockups: aircraft construction of. Washington, DC: May 13, 1969.

DoD, MIL-M-18828A, Mockups; construction of for target drones and guided missiles. Washington, DC: October 1, 1968.

DoD, MIL-STD-1472B, Human engineering design criteria for military systems, equipment, and facilities. Washington, DC: December 31, 1974.

Geer, C. User's guide for the test and evaluation sections of MIL-H-46855. Draft Report D194-10006-1, Boeing Aerospace Company, April 1977.

Helm, W. First interim report, human factors test and evaluation, function description inventory as a test and evaluation tool development and initial validation study, Report #SY-77R-75, Patuxent River, MD: Naval Air Test Center, 23 September 1975.

Helm, W. Third interim report, human factors evaluation of model P-3C update I airplane, Report #SY-122R-75, Patuxent River, MD: Naval Air Test Center, 12 February 1976. (a)

Helm, W. Fourth interim report function descriptive inventory as a human factors test and evaluation tool: An empirical validation study, Report #SY-127R-76, Patuxent River, MD: Naval Air Test Center, 30 July 1976. (b)

Holshouser, E. L. Guide to human factors engineering general purpose test planning (GPTP). Point Mugu, CA: Pacific Missile Test Center, TP 77-14, June 1977.

Ketcham-Wiedl, M. A. and Bittner, A. C. Anthropometric accommodation of a female population in a workplace designed to male standards. Point Mugu, CA: Pacific Missile Test Center, TP 76-3, 1977.

Malone, T. B. and Shenk, S. W. Human factors test and evaluation manual (HFTEMAN), Volume I: Data Guide, Volume II: Support Data, Volume III: Methods and procedures. TP 76-11A, B, C. Point Mugu, CA: Pacific Missile Test Center, April 1976.

McCauley, M. E. and Kennedy, R. S. Recommended human exposure limits for very low frequency vibration. TP 76-36, Point Mugu, CA: Pacific Missile Test Center, 29 September 1976.

Moroney, W. F. and Smith, M. J. Empirical reduction in potential user population as the result of improved multivariate anthropometric limits. NAMRL-1164, Pensacola, FL: Naval Aerospace Medical Research Laboratory, 1972.

NAVMATINST 3900.9, Human Factors. Naval Material Command, Washington, DC: September 29, 1970.

O'Connor, M. and Buede, D. Report on the feasibility of the application of decision analytic techniques to the test and evaluation phase of the acquisition of a major air system. Technical Report, Decisions and Design, Inc., April 1977.

OPNAVINST 3960.10, Test and Evaluation. Chief of Naval Operations, Washington, DC: October 22, 1975.

Schiflett, S. Operator workload: An annotated bibliography. Technical Report, Patuxent River, MD: Naval Air Test Center, SY-257R-76, December 1976.

ABOUT THE AUTHORS

Lieutenant Commander William F. Moroney, a Navy Aerospace Experimental Psychologist, is presently head of the Human Factors Engineering Branch of the Pacific Missile Test Center, Point Mugu, California. After receiving his Ph.D. from St. Johns University, he entered the Navy and has served tours at the Naval Aerospace Medical Institute and the Naval Aerospace Medical Research Laboratory in Pensacola, Florida. He has worked in the areas of selection and training, motion sickness, and aviation safety. He has several publications in the area of anthropometry, particularly as it relates to cockpit design. He has participated in the development and evaluation of several airborne weapon systems and is presently involved in the development of a light emitting diode (LED) helmet-mounted display and a display system to aid pilots in maximizing the capabilities of their aircraft. He is a member of the Human Factors Society, the American Psychological Association, and the Association of Aviation Psychologists.

Lieutenant Wade R. Helm is a Navy Aerospace Experimental Psychologist. He received his B. A. and M. B. A. from the University of New Mexico before entering the service, where he spent his first 5 years attached to operational units as a Naval Flight Officer and the last 5 years engaged in human factors research. He has served tours of duty with the Naval Aerospace Medical Research Laboratory, the Naval Air Test Center, and the Pacific Missile Test Center. His research efforts have included developing assessment methodology for the Naval Flight Officer Training curriculum, comparing adjectival and nonadjectival rating scale techniques in assessing psychomotor performance, determining the effects of verbal interference on psychomotor performance, developing a function analytic methodology for assessing system performance and workload, conducting comparative analysis of engineering design deficiencies on Navy aircraft, determining operator information processing capacities under complex task conditions, and applying multiattribute utility theory concepts to system test and evaluation. LT Helm's system experience include the S-3A, P-3B, P-3C, E-2C, EA-6B, A-4, EC-121, and F-18 aircraft. He is a member of the Human Factors Society and the Association of Aviation Psychologists.

THE CHARACTERISTICS OF NAVAL PERSONNEL AND PERSONNEL PERFORMANCE

Stanley A. Horowitz
Allan Sherman, LCDR, USN
Center For Naval Analyses
1401 Wilson Boulevard
Arlington, Virginia

ABSTRACT

The productivity of enlisted personnel aboard ships is measured as a function of their personal characteristics. Ship readiness, as measured by the material condition of shipboard equipment, depends on the size and composition of a ship's crew, the complexity of equipment, and other factors. The productivity of enlisted personnel varies systematically with high school graduation, entry test scores, paygrade, experience, Navy training, race, and marital status. The importance of particular factors varies by occupation. More complex equipment is in worse condition and requires higher quality personnel. Ship age and overhaul frequency also affect material condition. Implications are drawn for policies regarding recruitment, retention, manning, rotation, and pay.

INTRODUCTION

The efficiency of Navy personnel policies can only be judged by the contribution of personnel to the effectiveness of the fleet. This contribution is very elusive. Thus, little is known about the relative value of personnel who differ in such characteristics as education, experience, mental ability, and training in the Navy.

Proper allocation of Navy personnel requires that variations in productivity among individuals reflect variations in their cost. Thus, knowledge of how personnel differences are likely to contribute to effectiveness differences is necessary for rigorous analysis of Navy decisions regarding the level of manning, recruitment, assignment, rotation, and pay. Currently, these decisions usually reflect reasonable assumptions about what kinds of people are most suitable for what jobs.

This paper is an effort to improve personnel management and fleet readiness by focusing on the contribution of shipboard personnel to the material condition of equipment. If we are successful in attributing variations among ships in the level of maintenance to differences in crew members responsible for maintenance, we will have made an important step toward more informed analysis of defense manpower issues.

The study addresses a wide range of questions. Among the main ones are:

1. How valuable are different kinds of enlisted personnel in various maintenance occupations?

2. How could personnel policies be changed to improve the material condition of the fleet?

Although we focus primarily on personnel-related determinants of shipboard material condition, other questions are also dealt with in order to comprehensively examine the material condition of ships:

1. What is the contribution of more frequent overhauls to material condition?

2. How much worse is the condition of older ships?

3. How does equipment complexity affect material condition?

And, a related question: Are high quality enlisted personnel more valuable in dealing with more complex equipment?

The answers to these questions indicate that fleet material condition can be improved by revised personnel policies.

We found that the productivity of enlisted personnel is a function of their characteristics. In general, men in higher paygrades and men with more experience are more productive. High school graduation and entry test scores often predict performance. Training received in the Navy often enhances productivity. Older ships are in worse material condition, and lengthening the overhaul cycle degrades material condition.

The precise nature of the relationship between individual characteristics and productivity varies widely across enlisted occupations (or ratings). It also depends on the complexity of the equipment being maintained. Not only is complex equipment in worse condition, it requires more skilled men to maintain it. On the other hand, simpler equipment was found to benefit more from larger crews.

A MODEL OF THE MATERIAL CONDITION OF SHIPS

The amount of time that equipment fails to function in a specified time period can be expected to depend on the kind of equipment, the age of the ship, length of time since the ship was last overhauled, and manning. We use regression analysis to estimate the relationship between downtime due to shipboard equipment failures and its hypothesized determinants.

We have confined our examination to cruisers and destroyers: 40 destroyers (DDs), 18 guided missile destroyers (DDGs), 17 frigates (FFs), 4 guided missile frigates (FFGs) and 12 cruisers (CGs). These 91 ships are all the active ships of these types that underwent overhauls in fiscal years 1972, 1973, and 1974. To be sure that we were looking at comparable periods on all the ships, the entire period from one overhaul to a ship's next overhaul was considered.[1]

---

[1]The data we used on equipment failure were not available before 1970. Thus, we weren't able to look at the entire inter-overhaul period for some of the ships. At least 18 months of data were available for all the ships. We assume that the material condition of a ship is not a major factor in determining when it is overhauled.

Whenever a ship suffers an equipment failure that degrades its operational capability, it must file a casualty report (CASREPT). We have used CASREPT information to derive measures of maintenance effectiveness.[2] CASREPT downtime per month is our key measure of shipboard material condition.[3] CASREPT downtime is the number of casualties a ship had multiplied by the average time CASREPTs on that ship took to be fixed. CASREPT downtime per month is proportional to the average number of CASREPTs outstanding.

Rather than study the determinants of CASREPT downtime for entire ships, we concentrated on several subsystems. These subsystems were chosen because they are common to a large number of cruisers and destroyers, and are maintained by men in a small number of ratings. The subsystems are boilers, engines, gun systems, missile systems, antisubmarine warfare (ASW) systems, and sonars. Table 1 shows the ratings of the personnel who are responsible for the maintenance of these subsystems.

Table 1

Subsystems Studied

Subsystem	Associated Rating
Boilers	Boiler Technician (BT)
Engines	Machinist's Mate (MM)
Gun Systems	Fire Control Technician (FT) Gunner's Mate (GM)
Missile Systems	Fire Control Technician (FT) Gunner's Mate (GM)
ASW Systems	Gunner's Mate (GM) Sonar Technician (ST) Torpedoman's Mate (TM)
Sonars	Sonar Technician (ST)

As the table shows, the same ratings are sometimes responsible for part of the maintenance of more than one subsystem. To properly match men and equipment, we allocated CASREPTs both by rating and by subsystem.[4]

---

[2]CASREPT information is kept on an automated file system at the Navy Fleet Material Support Office (FMSO) in Mechanicsburg, Pa.

[3]We also examined data on material condition derived from 3-M corrective maintenance reports, overhaul departure reports, and INSURV reports (reports of the Board of Inspection and Survey).

[4]This allocation was accomplished by referring to the Equipment Identification Code (EIC) associated with each CASREPT.

The enlisted manning characteristics examined for our designated ratings are shown in the following list:[5] number of enlisted personnel; pre-Navy education; entry test scores; paygrade profile; length of service (LOS); time aboard this ship; time at sea; Navy schooling; specialized qualifications; race; marital status. The bulk of the personnel analysis in this paper relies on crew histories compiled from the Navy's Enlisted Master Record (EMR). To build these histories, we reviewed the records of the entire enlisted force for seven years between 1967 and 1975, and picked out the men on the 91 ships. We then developed aggregate statistics describing the characteristics of each crew by rating. This required weighting the characteristics of individuals by the fraction of the observation period they were assigned to the ship.

The level of CASREPT downtime should vary inversely with the number of enlisted personnel. Men with more pre-Navy education and higher entry test scores in relevant areas ought to do better maintenance. We expect more experienced men to be more productive than less experienced men, and men in higher paygrades to be more productive than men in lower paygrades. Since more experienced men are more likely to have higher rank, an analysis which focused only on rank, for example, would be unable to determine how much of the added productivity of senior men reflected selection of the best men for promotion and how much was merely the result of experience. By including both paygrade and LOS in the analysis, we will be able to disentangle the quality dimension of higher paygrade from the effect of experience. We will not assume that more experienced (or higher ranked) men continuously get better at their jobs. We will examine the possibility that, after a break-in period, junior men reach a higher level of proficiency beyond which they tend not to improve, or that further significant improvement only occurs after a considerable time.[7] Our estimates of the

---

[5]Data were also gathered on the age of enlisted men and on the number of officers aboard the ships, but these factors did not prove to be important.

[6]When characteristics changed during an individual's tour aboard one of the ships (e.g., LOS, paygrade), the change was taken into account. In many cases, we couldn't tell when men left the ships because they left the Navy and were not observed on subsequent EMRs. People who have been out of the Navy for 6 months are deleted from the EMR. Since there are 1- and 2-year gaps between the EMRs that we used, many men were dropped from the record before we observed them, it was necessary to approximate their departure dates from information on when they were likely to have left the Navy. In rare cases, information on personnel aboard DDs was taken from semiannual Bureau of Naval Personnel Enlisted Distribution and Verification Reports (BuPers Form 1080). Use of these data will be identified in context.

[7]Continuous linear and logarithmic forms were tried for the LOS variable. Then men were divided into eight LOS groups: under 1 year, 1-2 years, 2-3 years, 3-4 years, 4-5 years, 5-7 years, 7-10 years, and over 10 years. These classes were then aggregated to find the relationship that best predicted downtime. A similar aggregation procedure was used for paygrades.

relationships between rank, LOS and productivity will allow an alternative to the assumption that the pay of different kinds of enlisted men reflects differences in their productivity.[8]

Experience at sea may be more important in increasing the productivity of enlisted men than shore duty. We will examine whether men with more prior sea duty tend to have ships with less CASREPT downtime. We also will see whether ships with more stable crews, those whose men have been aboard longer, have less downtime. If either of these variables reflects higher productivity, the Navy's policy regarding sea-shore rotation will be open to question.

The completion of more Navy courses should lead to higher productivity, and thus to better maintenance.

The acquisition of certain advanced skills confers Navy Enlisted Classifications (NECs) on individuals. Some NECs can be gained only via school attendance; others can be earned on the job. We differentiated between these two types, and used the number of NECs of each type that men possessed as a measure of the extent of advanced training.

The impact of the race variable, the percent of the crew that is black, is not predictable, but its inclusion is nonetheless appropriate. If blacks receive lower quality educations, more blacks, holding educational attainment constant, may be associated with worse maintenance. If the Navy's entry tests discriminate against blacks, more blacks, holding test scores constant, may be associated with better maintenance. We hope to discover whether the Navy's use of high school graduations and of entry tests as guides to recruitment and assignment is equally appropriate for blacks and whites.

We are also unable to predict how marital status correlates with the productivity of enlisted men. Married men may be more stable and more serious workers, and hence more productive. On the other hand, some married men may dislike sea duty a great deal. This disaffection may make them less productive.

For each of nine groups (BT, MM, GM, FT, TM, ST, guns, missiles, ASW) we estimated a relationship for CASREPT downtime per month as a function of ship age, length of time between overhauls, equipment complexity, and the aforementioned crew characteristics.[9] Ships are the units of observation in the analysis.

---

[8]This assumption is used fairly widely. See, for instance, Formal and On-the-Job Training for Navy Enlisted Occupations, by R. Weiher and S. Horowitz, CNA Professional Paper 83, November 1971.

[9]We also examined the connection between operating tempo and material condition. No direct connection was found. In addition, the relative condition of ships based on the east and west coasts was examined. The west coast ships appeared to have less CASREPT downtime (they also steamed significantly more). Finally, using a procedure for looking at all our ratings simultaneously, we checked for whether there were systematic tendencies for some ships to be better than others in all areas. In some cases there were. Inclusion of these operating tempo, coast and ship variables did not have a large effect on the impact of other variables on CASREPT downtime, thus we have concentrated on the results of estimating the formulation described.

It was expected that newer ships would, other things being equal, have less CASREPT downtime.

A longer gap between overhauls should lead to more equipment downtime. If it does not, ships are being overhauled too frequently.

Ships vary to some extent in their equipment. Usually these differences correspond to ship type or class differences; sometimes they do not. Obviously, this may influence ships' maintenance histories. For instance, the 1200-pound boilers on some ships have had more problems than the older 600-pound type because of technical innovations in their design. In general, more complex equipment is expected to be down more often. Because of the differences between these two types of boilers, we allowed for the possibility that personnel contributions to the maintenance of boilers were different for ships with 600-pound plants and 1200-pound plants.[10] Equipment variations for the subsystems will be discussed along with the empirical results.

We estimated the relationship using ordinary least squares. As was noted earlier, the period of observation for the dependent variables was either the entire time between a ship's overhaul in FY 72, 73, or 74 and its previous overhaul, or as much of this period as possible (always at least 18 months before the more recent overhaul). For the explanatory variables, the entire inter-overhaul period was used. The condition of a piece of equipment depends not only on the care it is getting now, but also on the care it received in the past. This is why we've used such a long observation period, and why it seemed desirable to use a longer observation period for the explanatory variables than for downtime when the complete CASREPT data set was not available. We hoped to capture the long-run effects of variation in the determinants of maintenance effectiveness. The next section discusses the results of our estimations.

EMPIRICAL RESULTS

In this section the results of our estimations will be treated.[11] Due to extremely severe space constraints, only one of the relationships, that for boilers, will be discussed in detail. A summary of results will also be presented.[12] The explanatory variables differ across groups because variables that did not improve the prediction of CASREPT downtime per month were deleted.

---

[10]This was done by multiplying each personnel variable by both a 600-pound ship dummy and a dummy for ships with 1200-pound plants. The two variables were entered separately into the relationship being estimated. If this procedure did not improve the explanatory power, the results were discarded.

[11]Both linear and semi-logarithmic forms for the regressions were tested. The functional form that predicted best for a group is the one used.

[12]A more complete presentation of results appears in Personnel Performance and Ship Condition, CNS 1090, 31 March 1977, and is available from the authors upon request.

Boilers

For the most part, the ships have one of four kinds of propulsion plants. All of the DDs in the Forrest Sherman Class, all the DDGs, and all the CGs have 1200-pound per square inch (p.s.i.) main propulsion plants and two screws. The older DDs also have two screws, but 600 p.s.i. plants. The FF 1052 class has one screw and 1200 p.s.i. plants, while the FF 1040 (Garcia) class and FFGs have one screw and pressure-fired boilers.[13] Distinguishing among these kinds of systems proved to be very important in explaining the material condition of boilers as measured by CASREPT downtime.

Table 2 lists the CASREPT downtime for different kinds of plants. The more complicated 1200 p.s.i. plants obviously have more boiler trouble than 600 p.s.i. plants. Because boiler downtimes for the two types of one-screw plants were similar, they have been treated together in the rest of the analysis.

Table 2

CASREPT Downtime for Boilers

Ship classes or types	Number of ships	Kind of equipment	Average CASREPT downtime (hrs/mo) Boilers
CG, DDG, Forrest Sherman destroyers (except DD 933)	36	2 screws, 1200 p.s.i.	730[a]
FRAM destroyers	33	2 screws, 600 p.s.i.	218
FF 1040, FFG 1	11	1 screw, pressure fired	318
FF 1052	8	1 screw, 1200 p.s.i.	301

[a]730 is approximately the number of hours a month. This means that, on the average, these ships have one boiler CASREPT outstanding. Since they have two boilers, one is usually CASREPT-free. In any case, existence of a CASREPT does not necessarily imply complete inability to operate. Seventy-five percent of all CASREPT downtime is C-2, implying minor degradation of mission-essential equipment. If equipment is C-3 it is termed marginally ready. C-4 means not ready. In this study all three types of CASREPTs have been aggregated together.

---

[13]The 91 ships include one diesel-powered ship, one 600 p.s.i. ship with one screw, and one Forrest Sherman ship without automatic combustion control. All three ships were deleted from the BT analysis.

The predictive relationships estimated for equipment maintained by BTs are displayed in Table 4. The coefficients are the best estimates of the impact of a one-unit change in each of the explanatory variables on the average number of hours of boiler CASREPT downtime per month. These results are never in an unexpected direction and are often quite significant.

Ships with two-screw, 1200 p.s.i. plants had much more downtime than other ships.[14] Not only did equipment complexity affect material condition, it also affected the impact of the crew on material condition. Crew quality, as measured by entry test scores, paygrade, training, and length of service, seems to have mattered much more on 1200 p.s.i. ships, particularly those with two screws. We estimate that an increase of one percentage point in the average Shop Practices Test scores of BTs on two-screw, 1200 p.s.i. ships would lower CASREPT downtime by an average of 138 hours per month. There is also a very high payoff to having rated personnel. A one percentage point drop in the fraction of BTs who are unrated (E-3 or below) is associated with a drop of 25.19 hours in CASREPT downtime per month. Married BTs are less productive than single BTs on two-screw, 1200 p.s.i. ships. Perhaps they are less willing to put in the long hours the job requires. Training was important on one-screw ships, though not as important as on two screw, 1200 p.s.i. ships. If a quarter of the BTs attend one extra school, CASREPT downtime is estimated to fall by 72 hours a month (1/4 times 287) on the one-screw ships. Variations in crew size, on the other hand, appeared more important on 600 p.s.i. ships. Addition of an extra BT could be expected to decrease downtime by 71 hours per month.[15]

---

[14]The coefficient of 7924 does not mean that two-screw, 1200 p.s.i. ships have an average 7924 more hours of downtime a month than other ships. In cases like this, where different coefficients are estimated for different types of equipment, or where the characteristics that enter the predictive relationship differ by equipment type, one cannot look at the coefficient of an equipment-type dummy variable as reflecting the differential downtime of that kind of equipment. To derive the average difference in downtime per month by equipment type, one must use the entire relationship to estimate average downtimes for different kinds of equipment at reasonable values of the independent variables. A comparison of the numbers in the third column of Table 2 gives a good indication of the impact of equipment complexity on the material condition of boilers.

[15]The data underlying the crew size variable used here came from BuPers Form 1080. We gathered these data only for the DDs in the sample. Perhaps if we had had them for all 88 ships in this analysis, crew size would have appeared more important for the 1200 p.s.i. ships. (There were six 1200 p.s.i. DDs in this sample.) Using crew size data from the Enlisted Master Record, no crew size effect was found. Usually the EMR and 1080 form measures of crew size correlated quite highly (an average of .67). For BTs, the only rating for which 1080 form data were used, the correlation was only .48.

Table 3

Determinants of Material Condition for Boilers
(CASREPT downtime, hours per month)

Explanatory Variable	Coefficient	t-value
Personnel variables		
On two-screw, 1200 p.s.i. ships		
Average score on Shop Practices Tests	- 138	-3.34[a]
Percent of BTs who are E-3 or below	25.19	3.00[a]
Percent of BTs who are E-8 or above	- 34.06	-1.19
Percent of BTs with under one year in the Navy	35.65	2.50[b]
Average number of school-related NECs per BT	-1586	-4.26[a]
Percent of BTs who are single	- 23.20	-3.29[a]
On one-screw ships		
Average number of Navy schools attended by BTs	- 287	-1.87[c]
On two-screw, 600 p.s.i. ships		
Average number of BTs	- 71	-3.72[a]
On all ships		
Percent of BTs with under 10 years in the Navy	8.94	1.29
Nonpersonnel variables		
Equipment complexity		
Two-screw, 1200 p.s.i. plant	7924	3.60[a]
Logarithm of ship age (years)	515	3.22[a]
Constant	- 635	

Notes. Corrected R^2 = .52.
Degrees of freedom = 76.

[a]Significant at the 1 percent level.

[b]Significant at the 5 percent level.

[c]Significant at the 10 percent level.

These results do not mean that crew size makes no difference on 1200 p.s.i. ships or that Navy training makes no difference on 600 p.s.i. ships. They do mean that variations in these characteristics within the ranges observed in the fleet are not likely to make much difference.

Not surprisingly, we found that, other things being equal, older ships had significantly more boiler problems.

SUMMARY OF RESULTS

The material condition of shipboard equipment is affected by the complexity and age of the equipment, the length of time since it was overhauled, and the number and characteristics of the men who operate and maintain it. Crew characteristics that influence the productivity of enlisted men include high school graduation, entry test scores, race, marital status, length of service, paygrade, sea experience, and advanced training. Not all of these factors make a difference for all kinds of equipment, but in all cases some of them matter.

Our empirical results are summarized in Table 4. It displays the characteristics that we have found to influence the productivity of men in each of the six ratings we examined. It also shows other factors that affected the material condition of equipment handled by men in each of the ratings. An "X" signifies a relationship that was unexpected; a check means that it was not. A blank means that no relationship was found.[16]

Equipment complexity is an important factor in the condition of all kinds of equipment.

In all cases, men in higher paygrades are more productive than their juniors, even when length of service is held constant. Except for TMs, some measure of LOS related positively to productivity. For STs, sea duty is the only kind of experience that was found to increase productivity. Sea duty also is important in several other ratings.

Our results regarding paygrade and experience must be interpreted carefully. They mean that men who get promoted are more productive than men who do not under existing promotion policies. They do not mean that more men should be promoted. The mere act of promotion does not make men more valuable.

---

[16] In the rare cases where we found a relationship in a subsystem equation (guns, missiles, or ASW) that was not in the corresponding rating equation, it was assigned to the relevant rating in Table 4. Some of these estimated effects are more statistically reliable than others.

Table 4

Determinants of Personnel Productivity and Equipment Condition
As Measured by CASREPT Downtime

Crew characteristics or other determinants of material condition	BT	MM	GM	FT	TM	ST
Crew size	✓	✓		✓	✓	
High school graduation				✓		✓
Entry test scores	✓		✓	✓		
Paygrade	✓	✓	✓	✓	✓	✓
Length of service	✓	✓	✓	✓	X	
Sea experience {aboard prior ships		✓	✓			✓
Sea experience {aboard current ship		✓				
Training {number of schools attended	✓	X	X		✓	✓
Training {number of NECs attained	✓	✓	✓			
Marital status	✓					✓
Race				✓		
Ship age	✓	✓			✓	
Time between overhauls			✓	✓	✓	
Equipment complexity	✓	✓	✓	✓	✓	✓

In calculating productivity differences for men with different lengths of service one must take account of other factors that differ with LOS. For example, men who have been in the Navy 10 years are likely to be in higher paygrades than men who have been in 5 years. The probability of promotion and the estimated additional productivity of men in higher paygrades must be taken into account in comparing the value of men with different lengths of service.

FTs and STs are more productive when they are high school graduates. In other, less technical, ratings high school graduates were not estimated to be more productive than other men of the same paygrade and LOS. Entry test scores predict the performance of BTs, GMs, and FTs.

Variations in productivity reflected variations in training in all of our ratings except for FTs. Perhaps all FTs are so highly trained that variations do not matter much. When paygrade and LOS are held constant, however, additional school attendance helped MMs and GMs only if it led to attainment of an NEC. Interestingly, these were two ratings where sea experience was more valuable than shore duty in increasing men's productivity. Some of the value of training may have been picked up by paygrade variables. This will be the case if some men benefit from training and others do not, and if those who benefit are more likely to be promoted. We recommend extreme caution in using our results to draw negative conclusions about the value of training.

Single STs and BTs were estimated to be more productive than married men in those ratings.

Entry tests may discriminate against black FTs, who are more productive than expected on the basis of test scores and high school graduation. This effect was not found in other ratings.[17]

Older ships have more CASREPT downtime, particularly in engineering. Longer gaps between overhauls lead to more downtime in half of the ratings studied.

Table 4 misses some important facets of our results. Frequently, higher skill levels reflected in education, test scores, experience, or training increased productivity only when men handled relatively complex equipment. On the other hand, variations in crew size seemed to make the most difference on simpler ships.

CONCLUSIONS

We have answered most of the questions posed at the beginning of this paper. We have estimated the relative value of different kinds of enlisted personnel in different occupations, and shown how material condition could be improved. We have quantified the effects of ship age, overhaul policy, and equipment complexity on the ability of ships to perform their missions.

Our results have implications for what policies should be followed to improve the management of enlisted personnel. In many cases, discovery of the precise nature of these implications requires calculation of the cheapest way to improve material condition. This, in turn, requires that our estimates of productivity differences be combined with estimates of differences in the cost of personnel with various levels of education, ability, experience, and training.

In other cases the policy implications of our results are apparent without future analysis:

1. Place a higher proportion of senior men and highly trained men on ships with complex equipment.

2. Pay more attention to the level of manning on ships with less complex equipment. We would not recommend manning cuts where we found no impact of crew size because maintenance is not the only task men have.

3. Do not screen men so carefully on the basis of high school graduation and entry test scores in ratings where these characteristics do not seem to increase productivity.

4. Try to get sonar technicians to spend more time at sea. Paying special sea pay selectively to certain ratings should be considered.

5. Although higher entry test scores do not always indicate higher productivity, they usually do not seem to discriminate against blacks. Fire control technicians are an exception. Perhaps blacks should be given waivers to become fire control technicians even if they do not quite meet the usual criteria.

---

[17] CNA Study 1039, Enlisted Selection Strategies, by R. F. Lockman, found that entry tests are relatively poor predictors of the success of blacks in electronics schools in the Navy. (p. 10)

6. The current Navy data system is better for measuring material condition than many people believe. We have found reasonable relationships using the data.

7. More attention should probably be paid to the maintenance implications of introducing complex new equipment.

8. The Navy's policy of paying married men more than single men should be re-examined. Currently housing allowances and other benefits (exchange privileges, medical care) favor married men. Wherever we found a difference in productivity between single and married men, it was the single men who were better.

We found that the correlates of individual productivity and of subsystem material condition vary widely from rating to rating and from subsystem to subsystem. We have actually estimated relationships that have merely been asserted in the past. This study is the first we know of to go beyond the assumption that the relative value of men with different paygrades and lengths of service is measured by the ratio of their salaries. We know of no other statistical evidence that encouraging continuation at sea is important (aside from the possibility of cutting out superfluous shore billets). Also, there are few other indications that overhauls really do improve the subsequent condition of ships, and some work that calls the assumption into question.

By concentrating on CASREPT downtime as the measure of the condition of shipboard equipment, we have derived estimates that are relevant primarily for predicting changes in CASREPT downtime. Such changes may not correlate with other measures of material condition or operational capability, although they are correlated with both inspection results and records of 3-M corrective maintenance actions. In any case, CASREPTs are probably the best available information on ships' inability to perform their missions.

We feel strongly that efficient operation of the Navy requires quantitative links between the inputs that the Navy buys and the performance it delivers. This paper is one of the first such links for ship operations.

PERFORMANCE MEASUREMENT IN CIVILIAN ORGANIZATIONS: APPLICATIONS TO THE MILITARY SETTING

Mark S. Sanders Ph.D
California State University, Northridge

ABSTRACT

A comparison is made of performance appraisal as practiced in the military and in the civilian sector. The military is far ahead of civilian performance appraisal in many areas; however, in the area of assessment of management skills, in particular through the use of the "assessment center" concept, industry leads the military. The assessment center technique is discussed in detail. It is suggested as an approach which could be adapted profitably by the military

The question that was posed to me, and to which this paper is addressed, asked: "What can the military learn from performance measurement research and practice in civilian settings?" The answer, unfortunately, is "not a hellava lot." In fact, the military is far ahead of civilian performance appraisal in such areas as automated performance measurement, skilled performance assessment, and team/crew performance. One area, however, will be discussed, relating to the assessment of management skills, which is being widely and successfully applied in the civilian sector and could be adapted with utility in the military.

There are several reasons why civilian performance appraisal is, for the most part, handicapped in comparison with that found in the military. For one thing, government has had a headstart over industry in the area of performance appraisal. Emperors of the Wei Dynasty (221-265 AD), for example, were aided by "Imperial Raters" who appraised the performance of the members of the official family (Whisler and Harper, 1962). Despite this early beginning, however, it was not until the 1800s that government in the United States started appraising performance. Industry, on the other hand, didn't really get around to it until World War I.

More important than a slow start, however, has been the attitude among business and industrial leaders concerning the value of human resources. It is interesting that business and industry spend inordinate amounts of time and devote large numbers of people to inventory their capital resources (money, raw materials, machinery) yet are not willing to devote the same time and energy to inventory their human resources. Spriegel (1962) surveyed 567 companies and found that 256 had discontinued appraisal of executives and 184 companies had discontinued appraisal of foremen and lower level personnel. The most frequent reason given for dropping a program was that the time required for appraisal became excessive.

Coupled with this attitude toward human resources is the basic profit motive of most businesses and industries. What little performance appraisal exists is often directed toward the "bottom line." Simple measures of quantity and quality

PRECEDING PAGE BLANK-NOT FILMED

of production have become the metric for evaluation. Other dimensions of performance tend to be ignored and measurement techniques remain somewhat simplistic and primitive.

Another force in the civilian sector which has attenuated progress on performance appraisal has been the unions. Unions have stressed seniority as a determinant of promotion and pay. Performance, beyond minimum requirements, has taken a back seat in personnel actions. In essence, the incentive to management to invest in elaborate or sophisticated performance appraisal of rank and file personnel has been eroded to a large extent by this stress on seniority.

It is for these reasons that civilian performance appraisal has in general lagged behind or, at best, kept pace with military performance appraisal.

The Civilian Performance Review System

The annual or semi-annual performance reviews used extensively in both the military and civilian sectors are very similar and often suffer from the same shortcomings. A thumbnail sketch of the typical civilian performance review process will aid in identifying its weaknesses and may suggest parallels in military systems which need attention. Three studies have surveyed civilian performance appraisal systems: Spriegel (1962) surveyed 567 companies; Zawacki and Taylor (1976) surveyed 46 of the largest U. S. corporations; and Holley, Feild, and Barnett (1976) surveyed 39 organizations. These studies will serve as the basis for statements concerning common practices.

The aims or purposes of a performance appraisal system can be grouped into two broad categories: employee development and administrative action (Sanders and Peay, 1972). Appraisal programs are about evenly split between the two purposes, with most programs having multiple purposes (Holley et al., 1976; Spriegel, 1962). One rule, endorsed by virtually all writers in the field, is to limit the purposes of the appraisal program (Sanders and Peay, 1972). Often the kind of information needed for administrative action is counterproductive to employee development. Zawacki and Taylor (1976) report that 58% of the companies in their survey include the topic of pay in their discussions of appraisal with employees. This, in spite of the fact that trying to counsel an employee when he knows his salary (or promotion) hangs on a favorable evaluation, will cause him to become defensive and blame everyone and everything besides himself for his shortcomings (Meyer, Kay, and French, 1965).

Generally, the first level supervisor rates employees. There is evidence that immediate supervisors' ratings are more valid (Whitla and Tirrell, 1953) and are closer to the way the employees feel about themselves (Prien and Liske, 1962) than ratings of higher level supervisors. Generally, raters are not adequately trained. Often they are given little more than a group meeting explaining the program and a rating manual to read (Spriegel, 1962). Yet studies demonstrate that training increases both the validity (Bittner, 1948) and reliability (Stockford and Bissel, 1949) of ratings. A good training program may require several training sessions and workshops (Bittner, 1948).

The most common technique for obtaining performance appraisals is with the use of a numerical rating scale (Holley et al., 1976). Often global factors such as "quality of work," "quantity of work," "initiative" or "dependability" are rated (Holley et al., 1976). This is usually done without due consideration to

such problems as contamination, leniency, low reliability, or poor distinguishability of the traits. The majority of companies rate from 5 to 14 different traits, with some companies rating over 50, despite the prevalence of halo effects when so many traits are rated. It is extremely difficult for raters to distinguish more than five traits, with anything additional becoming redundant (Sanders and Peay, 1972).

Performance appraisal ratings are most often carried out on an annual basis. Most authorities are now suggesting at least semi-annual ratings should be done. It is interesting to compare the situation today to that in 1922 when Patterson recommended a three month interval between ratings rather than monthly.

As can be seen, little has occurred in civilian systems that is much different from what is already going on in military performance review systems. There is one appraisal paradigm, however, that seems to be unique to non-military organizations. This is the concept of assessment centers.

Definition of Assessment Centers

Finkle (1976) defines assessment centers as "a group-oriented, standardized series of activities which provide a basis for judgments or predictions of human behaviors believed or known to be relevant to work performed in an organizational setting."

According to Finkle (1976), the following four characteristics set assessment centers apart from previous managerial assessment approaches:

1. They operate with fixed sized groups of assessees.

2. They use several assessors serving in a nontraditional assessment role.

3. They employ multiple methods of assessment with strong emphasis on situational exercises.

4. They engender relatively favorable reactions from the assessees in the organizations in which they have been established.

Each individual element, itself, is not unique. It is the combination of them that sets assessment centers apart from traditional approaches of management appraisal.

In an effort to further contrast assessment centers to other assessment techniques, it will be of use to indicate what an assessment center is not. The following activities do not constitute an assessment center (Kraut, 1976):

1. Panel interviews or a series of sequential interviews as the sole technique.

2. Reliance on a specific technique (regardless of whether a simulation or not) as the sole basis of evaluation.

3. Using only a test battery composed of a number of pencil-and-paper measures.

4. Single assessor assessment (measurement by one individual using a variety of techniques).

5. Use of several simulations with more than one assessor where there is no polling of data (that is, each assessor prepares a report on performance in an exercise, and individual unintegrated reports are used as the final product).

History of Assessment Centers

Although the title of this paper is "Civilian Applications to Military..," it is interesting that the concept of assessment centers actually had its start in the military. The approach was first used by the Germans, then by the British in 1942, and finally by the United States Office of Strategic Services (OSS) in 1943 (OSS, 1948). Its principal use was in selecting intelligence agents for service during World War II. The OSS staff had taken seriously its obligation to check the validity of their operations, but for a number of reasons, all of which are spelled out in their book, Assessment of Men (OSS, 1948), they were unable to do so very effectively. Nor did other immediate post-war assessment programs fare very much better (MacKinnon, 1975), and, prematurely, as it turned out, Cronback (1955, 1956) declared that the OSS style of assessment had failed.

Cronback could not have known that one year later, 1956, Douglas Bray would revitalize the assessment method in the business world by beginning probably one of the most ambitious longitudinal research programs ever undertaken. Its purpose was to assess developmental patterns of beginning managers at AT&T (Bray, Campbell, and Grant, 1974).

The first assessment center in American industry was established then, not for operational goals of selection and placement, but for research purposes.

Growth of Assessment Centers

The first assessment center program established for operational purposes was instituted at Michigan Bell Telephone Company in 1958 (Michigan Bell Telephone Company, 1960). It was designed primarily as an aid to the line organization in the selection of high potential employees for managerial positions. The Michigan Bell center has been in continuous operation ever since its start and assesses approximately 600 men and women annually (Huck, 1973).

The concept quickly spread to other Bell System Companies, where today over 10,000 employees are evaluated each year (Huck, 1973). Other companies and government agencies began instituting assessment centers, including: Standard Oil Company (Ohio), Sears, J. C. Penney, IBM, General Electric, Minnesota Mining and Manufacturing Company, Peace Corps, Internal Revenue Service, and Oak Ridge Atomic Energy Facility. It was estimated that, today, over 1000 organizations use assessment centers (MacKinnon, 1975). The method is not only used in the United States, but in Canada, Australia, Japan, Brazil, and South Africa (Huck and Bray, 1976).

In 1973, Bender surveyed 32 organizations and found that their centers were in operation on the average of 2.5 years (range 1 to 15 years). Hence, although there are a substantial number of centers in operation, most are relatively new with few in operation longer than five years.

Principle Uses of Assessment Centers

The first operational assessment centers were developed to assess managerial potential among non-managerial personnel. This still remains the principal goal

of most assessment centers. In more recent years, however, the utility of assessment centers as a training device for personal development and growth has been recognized (Kraut, 1976; Anundsen, 1975).

Principal Features of Assessment Centers

The only attempt to document variations among operating assessment centers was done by Bender in 1973. Bender received questionnaires concerning 34 assessment center programs. Much of the data to be discussed in this section comes from that survey.

Length. Assessment centers vary in length from 1 to 6 days, with the modal lengths being 2 and 5 days (mean equal to 3.7 days). Most centers are carried out off the job, most often in a hotel/motel facility with assessees and assessors remaining at the location throughout the period (Bender, 1973). Interestingly, Moses (1973) reports that a one-day center yielded basically the same evaluations as a more expanded, lengthy, multi-day operation.

Activities. Bender (1973) reports that centers use from 4 to 40 different evaluation exercises with the majority of programs (71%) using 4 to 7 different exercises. The most frequently used evaluation devices were:

1.	In-basket	91%
2.	Leaderless group discussions	91%
3.	Business game exercises	88%
4.	In-depth background interviews	65%
5.	Psychological tests	59%

Each of these deserves a few words of explanation.

The in-basket exercise requires an assessee to sort and respond in writing to an accumulation of mail, reports, notes, etc., which might be left in the "in-basket" of a manager. Often the assessee is interviewed as to reasons for his actions.

Leaderless group discussions can be of either a cooperative or a competitive nature. The range of problems is limitless and sometimes reflects specific demands of the organization conducting the center. Most often the discussions involve such things as whether a hypothetical company should sell out, who among a group of hypothetical people should be promoted, how a disciplinary problem should be handled, or how money should be allocated.

Business game exercises may involve presentation of written reports or oral presentations. J. C. Penney (Byham and Pentecost, 1970), in an assessment center for product services managers, used a job relevant task. They had assessors, acting as irate customers, call the candidate and make several unreasonable demands. The assessee's ability to handle the situation was evaluated.

The in-depth background interviews, when used, usually cover an assessee's personal history, work history, and the history of his goals and values (MacKinnon, 1975).

Psychometric tests of mental ability, interests, values, and personality are administered in some centers, but typically the scores are not used in making decisions or recommendations. Rather, they are used as a check upon decisions or recommendations already agreed upon, or sometimes are retained for later research (MacKinnon, 1975).

In addition to these devices, peer, self, and sociometric ratings are often used in assessment centers.

The choice of assessment or evaluation devices is not a grab-bag process. Initially, the required skills and abilities for the particular level of job for which assessment is being made are determined. The selection of devices is then made to ensure that the important skills and abilities can be evaluated.

Candidates. Candidates are usually nominated by their supervisors and to a far lesser extent are permitted to nominate themselves (Bender, 1973). Usually the number of assessees evaluated at one time is six or a multiple of six (Bender, 1973), as this seems like an optimum number for group discussion exercises.

Assessors. Often managers are used as assessors. They are usually two or three levels above the assessees (Bender, 1973) and may or may not be assisted by professional psychologists. The ratio of assessors to assessees is usually 1:2 or 1:3 (MacKinnon, 1975). This ratio helps ensure in-depth analysis of each assessee. Byham and Pentecost (1970) believe that managers serving as assessors are more accurate in their judgments than when they conduct regular performance appraisals with their subordinates, for several reasons: (1) Not personally knowing the assessment center candidate, the assessor is unbiased and uninvolved emotionally. (2) The assessor can give full attention to observing behavior in an assessment center. (3) The specific behaviors that are to be evaluated have been identified and the assessor has been trained to observe them.

The amount of training given assessors varies from as little as five hours to as much as 15 days. The average number of days allotted to training is 4 (Bender, 1973). Minimum training requirements for assessors have been recommended by the Task Force on Development of Assessment Center Standards (reproduced in Kraut, 1976).

Some programs use a fixed group of managers as staff for several months, but often programs use managers for only one program. Arguments for frequent change stress the value of the assessment center experience for the assessors. The assessor/manager becomes more sophisticated in making human judgments and gains firsthand experience with the process, which helps to ensure its continued acceptance among managers. Arguments for less frequent change of managers as assessors concern the importance of greater stability in the program and the lower costs of training (Finkle, 1976).

Traits. Assessment is always of multiple traits. Finkle (1976) reports the traits rated from five assessment centers. The number varies from over twenty for AT&T to ten for the IRS. Bender (1973) lists the 26 parameters most frequently reported in his survey, but warns that an additional 40 were also listed. The following are the top ten traits as measured by the number of programs indicating evaluation (Bender, 1973):

1. Oral communication skills
2. Leadership
3. Organization and Planning
4. Decision-making skills
5. Problem analysis
6. Resistence to stress
7. Written communication skills
8. Energy
9. Use of delegation
10. Behavioral flexibility

The Report. The ratings and observations of each assessor are combined into a final narrative report about each assessee. Some programs (for example, Sohio) limit written reports to documentation of judgments, data and opinions developed by and/or agreed to by the entire assessment staff. The emphasis is on requiring the staff to offer only statements about how the assessee may be expected to behave in the future. More typically, however, assessment reports such as from J. C. Penney or Bell system programs, offer full elaboration on the observed situational exercise behavior as well as on the background of the assessee (Finkle, 1976). An example of a full-elaboration report can be found in an article by Byham (1970).

Usually these reports result in a global overall assessment of potential for each assessee. A common result is that 30-40% of the assessees are rated in the acceptable or outstanding categories, 40% are rated questionable, and 20-30% are rated unacceptable. This, despite the fact that the assessees were hand-picked by their supervisors as showing potential (Byham, 1970).

Most programs provide the assessees with feedback about their performance. It is always given orally and also may occasionally be given in writing (Bender, 1973). Kraut (1972, 1973) reports favorable reactions from assessees to their assessments.

Most assessment center programs also make the assessment known to top management and may or may not put the assessment report in the assessee's personnel file. Finkle (1976) warns against over use of the assessment center reports. He warns that such reports must be used in conjunction with on-the-job assessments before personnel recommendations can be made.

Validity of Assessment Centers

There have been numerous studies showing the validity of assessment centers for predicting mobility, promotion success, and on-the-job performance (see Huck, 1973; MacKinnon, 1975; or Finkle, 1976, for reviews of much of this literature). The studies generally show validity coefficients ranging from .30 to over .60, which is quite impressive. Most studies suffer from a problem of criterion contamination. That is, the assessment center report is made known to the management who in turn determines promotions and salary advances (two common criteria).

Two studies, both done at AT&T, however, do not suffer from such criterion contamination problems. Bray and Grant (1966) checked the assessment center's prediction concerning which candidates would reach middle management 5 to 8 years after the prediction was made. The point-biserial correlation between prediction and level achieved in management was .44 for the college men and .71 for the non-college portion of the sample. In the second try, Bray and Campbell (1968) correlated assessment center ratings with on-the-job performance of salesmen. They found that 100% of the candidates rated "more than acceptable" in the assessment center met the on-the-job performance standards for salesmen. The percentages meeting or exceeding the performance standards for each assessment center rating category were 60% of those rated "acceptable," 44% of those rated "less than acceptable," and 10% of those rated "unacceptable."

Cohen, Moses, and Byham (1974) compared the percentages of success for assessed and non-assessed groups from studies that made such comparisons. Almost without exception, the performance of people promoted into management position with assessment center recommendations was superior to that of people promoted without such a recommendation.

Kraut and Scott (1972), besides finding substantial correlations between assessment prediction made six years prior and promotions and demotions, also reported an interesting finding. They found that the proportion of low- and high-rated employees who left the company did not differ. This indicates that a low assessment center rating does not result in employees voluntarily terminating. This is encouraging, as candidates are evaluated only on managerial ability, not technical skills or current job performance.

In this age of equal employment opportunity, validity is often not sufficient. It is also important to show that an assessment technique is not biased with respect to women and minorities. Within the Bell System, women have been assessed since the early 1960s, initially in all-women groups, and starting in the later years of the decade, in integrated groups with men (Moses and Boehm, 1975). Bender (1973) found 73% of the companies he surveyed assessed females and 38% even used females as assessors. A full 85% of the companies assessed minorities.

Several studies have found, happily, that overall levels of performance of men and women in assessment centers do not differ (Huck, 1974; Moses, 1973). Assessment center methods, therefore, appear valid for the selection of women managers and do not result in the promotion of proportionately fewer women assessees (Moses and Boehm, 1975).

Huck and Bray (1976) also found that there was no differential validity of assessment center rating for black and white assessees.

In summary then, the assessment center approach seems to have substantial validity for long range predictions of managerial success across many different organizations and specific job situations. In addition, it seems effective for identifying qualified women and minority candidates without bias. As such, it is likely that assessment centers would have utility for the military, particularly in the selection of officers. In view of the expanding role of women in the military, the utility of the assessment center concept would be enhanced.

REFERENCES

Anundsen, K. An assessment center at work. Personnel, 1975 (Mar-Apr), 29-36.

Bender, J. What is "typical" of assessment centers. Personnel, 1973 (Jul-Aug).

Bittner, R. Developing an employee merit rating procedure. Personnel Psychology, 1948, 1, 403-432.

Bray, D. & Campbell, R. Selection of salesmen by means of an assessment center. Journal of Applied Psychology, 1968, 52(1), 36-41.

Bray, D., Campbell, R. & Grant, D. Formative years in business: A long-term AT&T study of managerial lives. New York: Wiley & Sons, 1974.

Bray, D. & Grant, D. The assessment center in the measurement of potential for business management. Psychological Monographs, 1966, 80, (17, Whole No. 625).

Byham, W. Assessment center for spotting future managers. Harvard Business Review, 1970, 48(4), 150-160.

Byham, W. & Pentecost, R. The assessment center: Identifying tomorrow's managers. Personnel, 1970 (Sep-Oct).

Cohen, B., Moses, J. & Byham W. The validity of assessment centers: A literature review. Monograph 11. Pittsburgh, PA: Development Dimensions Press, 1974.

Cronbach, L. Assessment, 1955. American Psychologist, 1955, 10, 419.

Cronbach, L. Assessment of individual differences. In P. R. Farnsworth and Q. McNemar (Eds.) Annual Review of Psychology (Vol. 7). Palo Alto, CA: Annual Reviews, 1956.

Finkle, R. Managerial assessment centers. In M. Dunnette. (Ed.) Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Holley W., Feild, H. & Barnett, N. Analyzing performance appraisal systems: An empirical study. Personnel Journal, 1976 (Sep), 457-463.

Huck, J. Assessment centers: A review of the external and internal validities. Personnel Psychology, 1973, 26, 191-212.

Huck, J. Determinants of assessment center ratings for white and black females and the relationship of these dimensions to subsequent performance effectiveness. Unpublished doctoral dissertation, Wayne State University, Detroit, MI, 1974.

Huck, J. & Bray, D. Management assessment center evaluations and subsequent job performance of white and black females. Personnel Psychology, 1976, 29, 13-30.

Kraut, A. A hard look at management assessment centers and their future. Personnel Journal, 1972, 51, 317-326.

Kraut, A. Management assessment centers in international organizations. Industrial Relations, 1973, 12 (No. 2).

Kraut, A. New frontiers for assessment centers. Personnel, 1976 (Jul-Aug), 30-38.

Kraut, A. & Scott, G. Validity of an operational management assessment program. Journal of Applied Psychology, 1972, 56, 124-129.

MacKinnon, D. An overview of assessment centers. TR No. 1, Center for Creative Leadership, 1975.

Meyer, H., Kay, E. & French, J. Split holes in performance appraisal. Harvard Business Review, 1965, 43(1), 45-51.

Michigan Bell Telephone, Personnel Relations Department, Personnel assessment program: A pilot study. 1960.

Moses, J. The development of an assessment center for the early identification of supervisory potential. Personnel Psychology, 1973.

Moses, J. & Boehm, V. Relationship of assessment center performance to management progress of women. Journal of Applied Psychology, 1975, 60, 527-529.

Office of Strategic Services (OSS): Assessment Staff. Assessment of men. New York: Rinehart, 1948.

Patterson, D. The Scott Company graphic rating scale. Journal of Personnel Research (Now: Personnel Journal), 1922-23, 1, 361-376.

Personnel assessment program: A pilot study. Michigan Bell Telephone, Personnel Relations Department, 1960.

Prien, E. & Liske, R. Assessment of higher level personnel. III. Rating criteria: A comparative analysis of supervisory ratings and incumbent self-ratings of job performance. Personnel Psychology, 1962, 15, 187-194.

Sanders, M. & Peay, J. Employee performance evaluation and review: A summary of the literature. RDTR No. 282, Naval Weapons Support Center, Crane, IN, 1972.

Spriegel, W. Company practices in appraisal of managerial performance. Personnel, 1962, 39(3), 77-83.

Stockford, L. & Bissel, H. Factors involved in establishing a merit rating scale. Personnel, 1949, 26, 94-116.

Whisler, T. & Harper, S. Performance appraisal. New York: Holt, Rienhart and Winson, 1962.

Whitla, D. & Tirrell, J. Validity of ratings of several levels of supervisors. Personnel Psychology, 1953, 6, 461-466.

Zawacki, R. & Taylor, R. A view of performance appraisal from organizations using it. Personnel Journal, 1976 (Jun), 290-299.

ABOUT THE AUTHOR

Mark Sanders is an Associate Professor, Psychology Department, California State University, Northridge. He is also a consultant to Canyon Research Group, Perceptronics, and the Naval Personnel Research & Development Center. He received his Ph.D. degree in 1971 from Purdue University. A member of the Human Factors Society since 1967, he is a member of the editorial board of *Human Factors* and the present chairman of the Membership Admissions Committee. He is author of the "Workbook for Human Factors in Engineering & Design". He has published numerous technical reports and journal articles as well as presented numerous works at professional meetings. He is a member of the American Psychological Association, Division 14 & 21.

PERFORMANCE MEASUREMENT SYSTEM ARCHITECTURE AND DATA PROCESSING LOADS

R. W. Obermayer
Navy Personnel Research and Development Center
San Diego, California

ABSTRACT

The point of view taken in this paper is that the performance Measurement System (PMS) is an information system that provides information for decision. The emphasis is placed on PMS processing loads and architecture, including treatment of some of the PMS parameters which may be manipulated during PMS design. An example PMS for flight research was used to show the types of difficulties which may be encountered during PMS design and operation.

The importance of the utility concept for PMS design is stressed: in short, is the desired information worth the costs involved? However, it is noted that processing load and schedule considerations may overshadow utility considerations during PMS design deliberations.

Several suggestions are made for the reduction of PMS loads: (1) attempt to implement only necessary and sufficient measurement; (2) reduce the volume of data through experimental design, through minimizing the number of parameters processed, and through minimizing the sampling rate; (3) expect errors and design the PMS to offer automatic detection and correction assistance; and (4) use a highly automated PMS to provide early transformation of raw data and quick interpretable results, as well as consideration of other types of automated assistance.

The PMS <u>is</u> the source of information for decision and more care than usual should be given to its design.

INTRODUCTION

The need for information leads to measurement, and it is important to keep in mind that the purpose of measurement is to provide information. This point of view is important for the development of appropriate measurement, for such development must begin with this question: What do you want to know? It is also important to maintain this point of view for the development of the Performance Measurement System (PMS), for the PMS is, in essence, an information system.

The PMS collects data and performs necessary transformations to provide information. For example, data may be collected by recording the indications of a measurement equipment panel meter; however, these readings may not be informative unless they are related to other different meter readings or are statistically averaged during a specified time. The PMS consists of all of the human and machine components which provide meaningful information, compiled into a comprehensive data base for analysis. The PMS is also a conceptual structure as diagrammed in Figure 1, which will be used in this paper to identify and analyze performance measurement difficulties in large data-collection efforts.

PRECEDING PAGE BLANK-NOT FILMED

The PMS for a large-scale system is often in itself a major, costly system. Large amounts of raw data are collected and processed into a form for decision. The PMS may bog down under the heavy data processing loads and ultimately produce little and/or misleading information. In short, the whole activity can be a waste if the PMS is not designed properly. Since the PMS is really not any different from other data gathering and processing systems, general data processing system design techniques for reliable information processing without overload are also appropriate to the PMS.

This paper discusses the design of the PMS in terms of architecture and loads to provide essential information in a cost-effective manner. The form of a PMS as addressed in this paper is the portion of Figure 1 within the dashed-line border. Note that the output is directed to the decision-makers and that the research scientist has the role of an interpreter to transform information finally to the proper decision form; this process is certainly important, but the related issues of such transformations are beyond the scope of this paper.

The input to the PMS is a sensor which may be a device or a human data collector. In either case, information is sensed and transmitted to the system, correctly or incorrectly, and perhaps with some degree of distortion. Since no sensor is totally reliable, some forms of data editing must exist. Frequently this is just a manual/visual scan, but in a system with a heavy volume of data there is a need for automatic assistance. As data are collected and transformed, an accumulative database is compiled. The raw data probably will also be recorded so that the database could be recreated (possibly with new measures as the result of new transformations). The data may be played back during this process; e.g., at an early stage for feedback to the subjects and/or review of the data for errors and for preliminary test-analyses.

Of course, there are many variations of the PMS shown in Figure 1, but the stages of sense, edit-correct, transform, update, playback-review, and analysis are usually found in any large PMS. The problem is to keep data flowing smoothly through such a system with acceptable accuracy and in such a way that the desired information is produced at the end of the pipeline. Bottlenecks can occur at any point in the process, and the design issues related to avoiding bottlenecks will be treated throughout this paper.

An Example

As an example of a complex performance measurement system, a study by Obermayer and Nicklas (1973), which was performed for the purpose of developing a measurement system, will be briefly described. The performance measurement system was expressly developed for use with a high-performance (Mach +3) research vehicle with a two-man crew. The primary emphasis was placed on the assessment of crew workload and on the development of a comprehensive PMS and set of measures for the many facets of workload exhibited during stressful flight.

The research vehicle provided an intensive workload situation with emergency or unexpected occurrences commonplace. In this regard, the unstart condition is worthy of special mention. At supersonic speeds, a series of shock waves occur in the engine air inlet duct. A suitable flow of air at proper pressure levels is necessary for engine operation. At the limits of engine performance, the shock waves may be forced out of the inlet, severely hindering the flow of air

to the engine, with a subsequent loss of thrust called an unstart. An unstart may be mild, may be severe enough to batter the crew in the cockpit, or may even cause pitch control problems which may lead to loss of control. It may be seen that the possibility of an unstart was of some concern to the flight crew, and, while some unstarts were planned or expected, numerous unexpected unstarts occurred.

Data Sources. Data were available in this environment from a number of sources, including (1) biomedical data from a cockpit recorder, (2) recordings of flight parameters in the form of pulse-code-modulation tapes (PCM) and/or information telemetered to ground recording stations, (3) communications recordings, (4) crew interviews, and (5) observations collected in the Mission Control Center during each flight.

A staggering amount of information was available from these sources. The biomed recordings included EKG, respiration rate, respiration volume, acceleration, and audio; all of which were recorded throughout the flight. There were four systems for flight recording, with 80 channels to each system, for a total of 320 channels; each channel was recorded at a rate of 200 samples per second, for a total of 64,000 data points each second. Additional data were available in the Mission Control Center in the form of flight path plots and strip-chart recordings, and also from the flight crew in the form of flight notes and subjective measurement.

Data Processing Procedure. An overview of the data processing procedure is shown in Figure 2. Biomed data were collected in analog form, requiring that the data be sampled and digitized, and specific parameters such as heart rate were calculated from these data. The basic physiological data thus provided were found to have significant errors which were, however, computer correctable. Desired measures were then computed from the corrected information. Voice communication recordings were tediously transcribed. Again, errors were found, especially in the timing of events from these tapes; however, the errors gradually accumulated throughout the tapes and were also computer correctable based on external accurate time information. Flight recording information was taken from three of the four system magnetic tapes and was transformed into computer-compatible form. This step was such an extensive process that it was only performed for specific time segments ordered in advance. Consequently, it was especially important during flight monitoring to identify the segments of time desired for analysis. The flight engineer's log and the audio recordings were also valuable during the data-ordering process. When the flight recordings were made available, careful manual and computer editing was again required to identify data losses, inoperative channels, and inappropriate time segments prior to computation of performance measures. When all computer-derived measures were collected, comparisons and correlations with subjective measurement were accomplished.

Measurement Computation and Analysis. A battery of measures was generated for the evaluation of alternative methods for the measurement of workload, including: (1) measures derived from EKG recordings, including heart rate and sinus arrhythmia measures; (2) information quantified from audio communications recordings; (3) subjective measures obtained from crew interviews and ratings and also selected commentary from flight communications recordings; (4) measures obtained by fitting mathematical models to pilot control data, quantifying the manner with which the pilot performs manual control of the vehicle; and (5) measures of system performance on the tasks required of the crew to perform the flight test mission.

Inferences about pilot performance and associated workload can be obtained using the above battery of measurements by making comparisons of performance (1) just before a flight test maneuver to that during the maneuver (also early in the maneuver to that later in the maneuver), (2) during normal maneuvers such as cruise, climb-descent, turn, etc., to performance during the same maneuvers with superimposed propulsion tests, and (3) during control of primary flight parameters as opposed to control of secondary, or lower priority, flight parameters.

General Criteria for PMS Design

The point of view taken in this paper, as expressed earlier, is that the PMS exists to provide information and that, ultimately, the information supports a decision. The process of measurement is supposed to provide the needed information; therefore, the effectiveness of the measurement should be evaluated in terms of the utility of the information for decision making.

In classical terms (Cronbach & Gleser, 1957; Wald, 1950; Girshick, 1954), the utility of measurement is determined by the following formula:

UTILITY = (Cost of erroneous decision) - (Cost of measurement).

Of course, the cost of erroneous decision is often difficult to assess, but it is important to note that often a satisfactory decision can be (or must be) made on less information than one would desire, and that measurement is a costly process; the possible avoidance of erroneous decision may simply not be worth the cost.

One might be tempted to present a list of measurement criteria consisting of various forms of validity (Obermayer, 1964) and reliability. However, validity is often a vague concept difficult to apply in practive, and the scientist may wish to substitute subjective estimates of comprehensiveness, accuracy, and reliability. To be on the safe side, various forms of redundant and similar measures may be included in the measurement set. However, one should first identify the decision to be made and then the minimal information satisfactory for the decision. The cost-effective measurement system may then be more readily defined if this general procedure is followed.

Processing Requirements. Performance measurement design will also be heavily influenced by the schedule governing the time at which measurement information is needed. A typical list of measurement system processing requirements may be:

1. Daily summary data (or before next major data collection).
2. On-the-spot briefing.
3. Weekly database update.
4. Early trial analyses.
5. Lead time for final summary analysis.

A schedule requirement exists because of the need to monitor the quality of measurement as it is collected; for example, the measurement should be examined

regularly (e.g., at the end of each day) to determine if part of the measurement system is malfunctioning. This way, the problem can be fixed before unacceptably large quantities of data are lost and cannot be recreated. Often, there is a requirement for briefing crews or providing feedback to subjects, making some information necessary near the end of each trial. Regular database updates are desirable to provide a broader view of the nature of the data collected, to permit thorough editing, and to make backup copies to prevent accidental loss of data. Early trial analyses are desirable for feedback on study trends, for modifications to procedure if absolutely necessary, and, as will be seen, to provide the possibility to reduce and streamline the PMS. Finally, of course, data processing of all measurement must be completed with sufficient lead time for final analysis and reporting. In short, this sort of processing requirement will provide specific demands on the performance of the PMS and may have a strong influence on the amount and type of information which can be made available through measurement.

Measurement Loads. The capability to process measurement loads on schedule is such an important design consideration that this paper mainly stresses methods for preserving measurement quality in the face of such loads. In the flight research example, performance parameters are generated at a rate of 64,000 datapoints per flight second; therefore, a selected minute of a flight (the basic unit used for measurement in the example) could require the processing of 3,840,000 data points. Of course, one must be more selective in the specific parameters actually needed for measurements. For the example, however, the minimal parameter set was about 32 parameters, resulting in the production of a potential 384,000 data points per flight minute, which would still be an overwhelming amount of data for a large number of flights averaging about 1½ hours (34,560,000 data points) per flight.

The Necessary and Sufficient Measure Set

A principle that appears to be commonly used in the generation of measurers for performance measurement is, "If it moves, measure it." This is, parameters which vary during a segment of performance are often considered important for measurement just because they are known to change, and knowledge of all types of variations in these parameters is considered important. In view of the high costs of measurement and the levels of measurement loads which develop, it may be more appropriate to follow the following procedure: (1) identify necessary information requirements, (2) identify candidate measures which impart needed information, (3) examine the utility of each measure, and (4) select only the needed measures from the set of candidates. The latter procedure is reflected in the flow diagram shown in Figure 3, and this is the logical point to start in the development of the PMS.

Generation of Candidate Measures. Some insight into the process of generating measures can be gained by examining the methodological framework which is used to conceive measurement. For this purpose a generalized model is presented in Figure 4 (adapted from Finley, Obermayer, Bertone, Meister & Muckler, 1969) in the hope of achieving some clarification.

What the figure says, in effect, is that for performance measurement in person-machine systems we must be concerned with three levels of measurement analysis: (1) system requirements and appropriate system performance measurement, (2) human

operator task analysis and the performance measures related to that level, and (3) basic behavioral dimensions involved in human task performance. In addition to the measurement indicated in Figure 4, consideration should also be given to psychophysiological performance measurement such as workload and the amount of energy expended while performing work. No attempt will be made (nor is it presently possible) to accomplish an integration of the person-machine performance dimensions; rather, the point to be made here is that this approach will generate a large (often overwhelmingly large) set of different, but interrelated, measures. In short, this is where many data processing load problems start.

Consider the task of measuring throughout a naval mission: the mission may be divided into maneuvers, the maneuvers into segments, and, within each segment, a number of tasks with a number of task dimensions may be measured. For each portion of the mission thus defined, system, task, and behavioral measures may be developed. With this background, it may be apparent that a large number of candidate measures may be defined. In fact, for a simple captive helicopter performing a series of common maneuvers, over 800 measures were generated (Vreuls, Obermayer, Goldstein & Lauber, 1973). Faced with the practical and conceptual difficulties encountered with this many measures, the desirability of achieving any possible reduction is evident.

The Necessary and Sufficient Measure Set. Given the system analysis, task analysis, and behavioral dimension analysis of Figure 4, one may attempt to choose specific measure forms. Normally, the result will be a large set of candidate measures. Many of these will be equivalent alternative forms; others may prove to provide unnecessary information or none at all. First, however, it can be asked if this measure set is sufficient; that is, are all the system, task, and behavioral phenomena accounted for in sufficient detail? Next, it can be asked if the resulting measure set is necessary; that is, are all the measures really needed and, in particular, needed for the specific decisions at hand? For measurement in a complex system environment, the quantities of measurement required and the difficulties in effecting measurement combine to require that all unnecessary and redundant measures be eliminated.

The question of sufficiency is one which depends largely on the effectiveness of the analytic techniques used and on the research which has been undertaken to understand the phenomena involved. The question of necessity is one which can be answered if operational criteria for necessity can be specified.

At least three general criteria may be defined to guide the reduction of the candidate measure set:

1. Dis ard those measures which provide the same information; that is, discard those that correlate highly.

2. Retain only those measures which are able to discriminate critical performance differences; that is, retain the measures which are able to discriminate (for example) between "good" and "bad" performers or students and instructors, and to discriminate performance changes during training. For the flight research example given earlier, measures which discriminate between tasks with different levels of workload can be determined through statistical methods. Nevertheless, selection of required measurement remains an important and largely unsolved issue, often more critical from a practical point of view than developing new forms of measurement.

3. Retain those measures which relate to or predict performance of interest; for example, those that are able to predict terminal training performance or deficiencies requiring special attention.

An intercorrelation analysis, correlating each measure with every other measure, provides a means for determining redundant measures (criterion 1). Multivariate statistics provide means for testing measures with respect to criteria 2 and 3. Given data from groups known to be different in some way, a multiple discriminant analysis yields the information to determine the measures which contribute to the discrimination. A canonical correlation analysis will allow determination of those measures which relate to (or correlate with, collectively) other measures (for example, measures taken early in training vis-a-vis measures taken late in training). Computer programs are available which can reduce the candidate measure set in a stepwise iterative fashion, allowing one to reduce the measure set to a manageable size with specific known properties of the measure set ultimately retained.

Although some analytical tools are available for measure selection, the problem is not totally resolved. Data must be collected from groups of subjects with known properties; consequently, two difficult tasks arise: (1) time-consuming, laborious data collection to perform measure selection as needed, and (2) the identification of subject groups with known characteristics (probably subjectively identified). Consequently, while some techniques for measure selection are available, additional developments are in order to increase the effectiveness and efficiency of measure selection.

Reducing Volume

Although it should be obvious, for the sake of completeness it should be mentioned in passing that the measurement load can be reduced if one just doesn't collect as many data points. For example, an efficient experimental design will not create the high measurement loads that an inefficient design will. Perhaps not as obvious are those cases where the same measurement can be based on fewer performance parameters, and when parameters need not be sampled as often.

For the flight research example, it was found that the parameters of interest were distributed over three of the four PCM system tapes recorded for each flight. This caused severe computer processing difficulties with long delays because such processing used all of the magnetic tape drives and prevented use of the machine by more than one user at a time. After some experience was gained, it was found that the pertinent parameters were recorded mostly on one tape, and that the others were either also recorded on the biomed tape or could be computed from the parameters on the one tape. After simplifying to a one-tape procedure, the measurement load decreased drastically: run time was reduced 50 percent; total delay was reduced from weeks to generally overnight.

Each parameter on the PCM tapes was recorded at a rate of 200 samples/second but could not be easily processed further at this sampling rate. An important sampling theorem (Blackman & Tukey, 1958) can be stated in terms of the bandwidth of the original parameter signal: the minimum sampling rate, with respect to reproducing the original frequency content from the digital signal, is twice the highest significant frequency. Briefly, to retain the original spectral qualities, the sampling rate must be numerically greater than twice the bandwidth. For example, if the

bandwidth (highest frequency) is 1 Hertz, the data must be sampled at a rate of 2 samples/second. For the bandwidths which are within human capabilities for manual control, this rule translates to a maximum sampling rate of about 10 samples/second. For the specific control signals recorded, and considering the bandwidths of the sensors used, a sampling rate of 5 samples/second was found to be acceptable. Sampling at a higher rate would only have increased measurement load, as no additional information could have been derived from signals sampled at a higher rate.

Designing for Error

The saying that "if something can go wrong, it will," suits the performance measurement system well. After reduction of the measure set, parameter set, and sampling set, the primary bottlenecks will probably be in the edit-correction loop of Figure 1.

All data must be examined visually because the unexpected often occurs, making full automation impossible at this point. However, if the data are stored in digital form on some computer medium, the error must be replaced through some use of the computer. Furthermore, some errors occur so regularly that a computer program must be prepared to detect them and to make the usual correction. On-line display methods can be developed so that human surveillance can be maintained, and these "automated" detection/correction procedures can be approved at each step while the operator maintains a lookout for any other anomalies.

It is quite important to detect problems early so that remedial action can be taken before it is too late. The need for daily examination of data (at worst, overnight processing of data for examination before the next day's data collection) should be given heavy emphasis to sustain the type of processing schedule listed under the section on Processing Requirements. For example, the biomed recordings for the flight research study lagged behind for about 6 weeks at one point. When they were finally examined, it was found that there had been a problem with the flight recorder, causing the loss of biomedical data for five critical flights. Other biomed difficulties which may be of interest for anticipating errors for PMS design are: (1) data were labelled with the wrong time, periods of data were missing, and a "minute" of data was sometimes longer or shorter than a minute; (2) the computer program measured exceptionally long and short heart beats occurring in pairs; and (3) beats were skipped, measured with zero duration.

Long delays were also encountered with the flight research telemetry data tapes; however, a significant part of the processing delay was reduced when the requirement was reduced to one tape instead of requiring the merger of three tapes. A number of aircraft parameters were found to be incorrect without further computations to correct for temperature, altitude, etc. The data tapes often were not as ordered; for example, the data were found to have different start/stop times, and sampling rates and the segments ordered were found to be combined or split into two parts. One data tape was overwritten by some user so that the data from that flight was not usable. The time of occurrence of flight events was sometimes not as reported, requiring careful monitoring of time in the mission control center and measurement of time from the audio tapes; even so, the time recorded on the tape was sometimes inexplicably different. A further difficulty was largely due to inadequate understanding of the nature of some tests; for example, some data collected to analyze crew control performances were found to have included no operator control since the system was in an automatic flight mode.

Moral: Design for the occurrence of errors in the PMS. A flexible, semiautomatic system should be designed to aid in the detection and correction of errors. Do not assume that any data are correct until examined carefully with the aid of computer tests.

The Automated Performance Measurement System

It should be noted that the flow is greatly reduced in Figure 1 at the point where measure transformations occur. Here, a number of parameters sampled many times per second over a period of time enter a computation which yields only one number. It may be seen, therefore, that a useful general rule for the reduction of measurement load is to transform early. Further, it is wise to record raw data only over those periods of time required for the measurement transformations; that is, do not record data for the full mission duration just to compute measures for selected minutes of performance. However, greater periods of time are often recorded to be sure to include the needed time period; these data are then computer-scanned to determine the point at which measurement transformation should start and stop. Additionally, it is desirable to produce directly interpretable results as early and as often as possible throughout a study to detect unexpected errors and to develop early indications of data trends.

An automated performance measurement system (or subsystem) can conceivably accomplish these ends: transformations can be calculated early, often as the data are being collected; the computer algorithms can detect stop and start times, avoiding excessive data recording (in fact, recording is only a backup for selected replay and processing); and directly interpretable results (the performance measures) can be displayed at the end of each mission, or possibly at the end of each mission segment.

A performance measurement subsystem of the type suggested by Knoop (1968) and Vreuls and Obermayer (1974) is presented in Figure 5. A digital computer is used to monitor sensed parameters during each maneuver comprising the man-machine system mission; different parameters may be monitored during each maneuver. The portions of the maneuver during which measurement is desired are called segments; when the segments are defined logically, the computer is able to determine when measuring should begin and end. During each segment a number of measures may be computed according to specified transformations of each parameter. The measures are subsequently stored for later analysis. The structure of the performance measurement subsystem can be fixed, and tables can be entered by the scientist to define start/stop conditions, parameters and sampling rates, and measure transformations.

The technology for performance measurement subsystems has been developed and tested. However, the specific design will depend upon the equipment configuration available, measurement information requirements, and constraints such as size, weight, power, heat, and cost.

While the basic structure of the performance measurement subsystem is shown in Figure 5, other associated functions should be included as indicated in Figure 1, such as data editing and analysis programs. Statistical analysis programs are, of course, needed to process the measures stored in the data base; and the need for data editing should not be overlooked under the incorrect assumption that errors do not occur in automatic computerized systems. Errors do occur frequently

in such systems for reasons such as described in the previous section. Because of the volume of data processed, the investigator requires computer assistance in both error detection and error correction.

A Procedure for Performance Measurement System Development

The previous sections of this paper present information which, it is believed, should be useful for the design of new PMS and for the redesign of the poorly performing PMS. Partially as a review, and partially to show the integration of this information into a design effort, the following paragraphs present four steps for PMS development. Of course, much of this procedure is relevant to the topic of performance measurement as a whole, and not just the systems for measuring.

Step 1. Analytical human-machine system study.

a. Perform analyses of system, mission and functions, analyses of human operator-maintainer tasks, and analyses of socio-psychological behavior.

b. Identify critical dimensions of the system, tasks, and behaviors.

c. Specify system, task, and behavioral measures.

As a first step in measurement development, as a general rule, the information implied by the general methodological model of Figure 4 must be made available. The information for measurement, as in Figure 4, requires the conduct of a series of analytical studies. We must define what we know about the system and the relationships between the subsystems to determine the criteria relevant to the performance of the defined mission. These must be translated into the terms of human tasks and human implications. From this, a candidate measurement set can be developed. As previously discussed, such an approach often leads to "overkill."

Step 2. Empirical Measurement System Study.

a. Collect sample data for candidate measures to permit statistical reduction to the necessary and sufficient measure set.

b. Collect sample data for determination of optimal sampling size and sampling rate.

c. Collect data on measurement system errors.

Step 2 is one which is generally omitted, but it is the step which, in practice, permits design approaching the optimum. Costly and time-consuming data collection efforts are involved; however, the benefits will include better measurement as well as improved PMS performance. Data are required to (1) reduce the candidate measures to those that are necessary and sufficient, (2) reduce data collection for the production of the selected measures, and (3) provide measurement experience (especially in regard to PMS errors) leading to optimal PMS design. These requirements may be met individually, but if a block of effort can be set aside for all three objectives, then a sequence of studies may be performed which satisfy all requirements more efficiently.

Step 3. Design the PMS.

a. Optimize data processing loads.

b. Design for error detection and correction.

c. Design for automation.

d. Collect data for the computation of measurement utility.

In short, one should specifically design the PMS, using principles such as those developed in this paper, and not just let the PMS grow "like topsy." This paper has been primarily concerned with this step.

Step 4. Test the PMS and Iterate. As a final step, since the previous steps provide no guarantee that the PMS will be totally satisfactory, the PMS must be tested under simulated or sample conditions to provide data on PMS performance. Given data on less than satisfactory PMS performance, a diagnosis and redesign can be attempted to again provide the needed PMS. Unfortunately, this type of iterative PMS design approach seems to be the best that can be offered with the current state of the art.

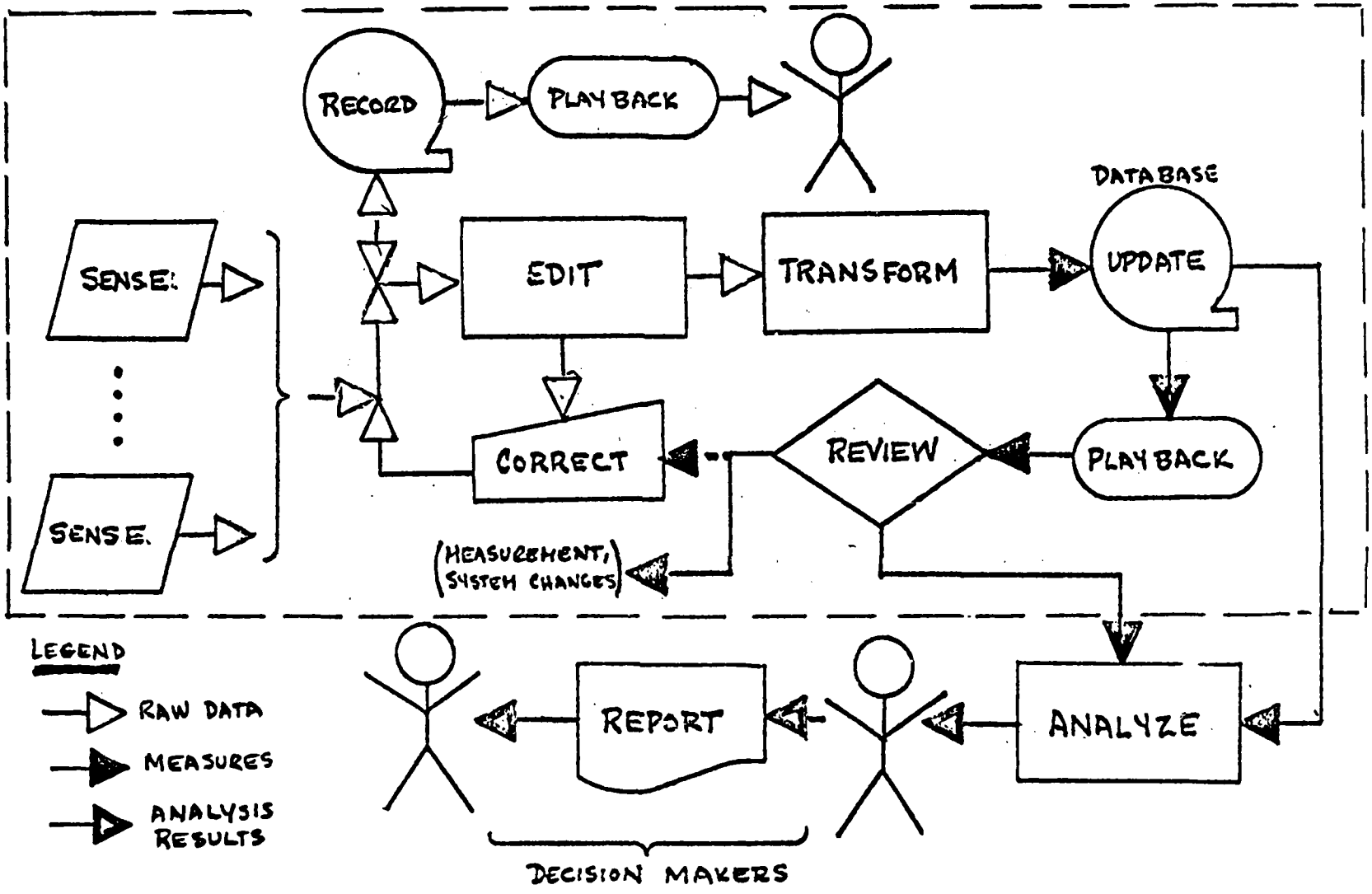


Figure 1. Performance Measurement System architecture.

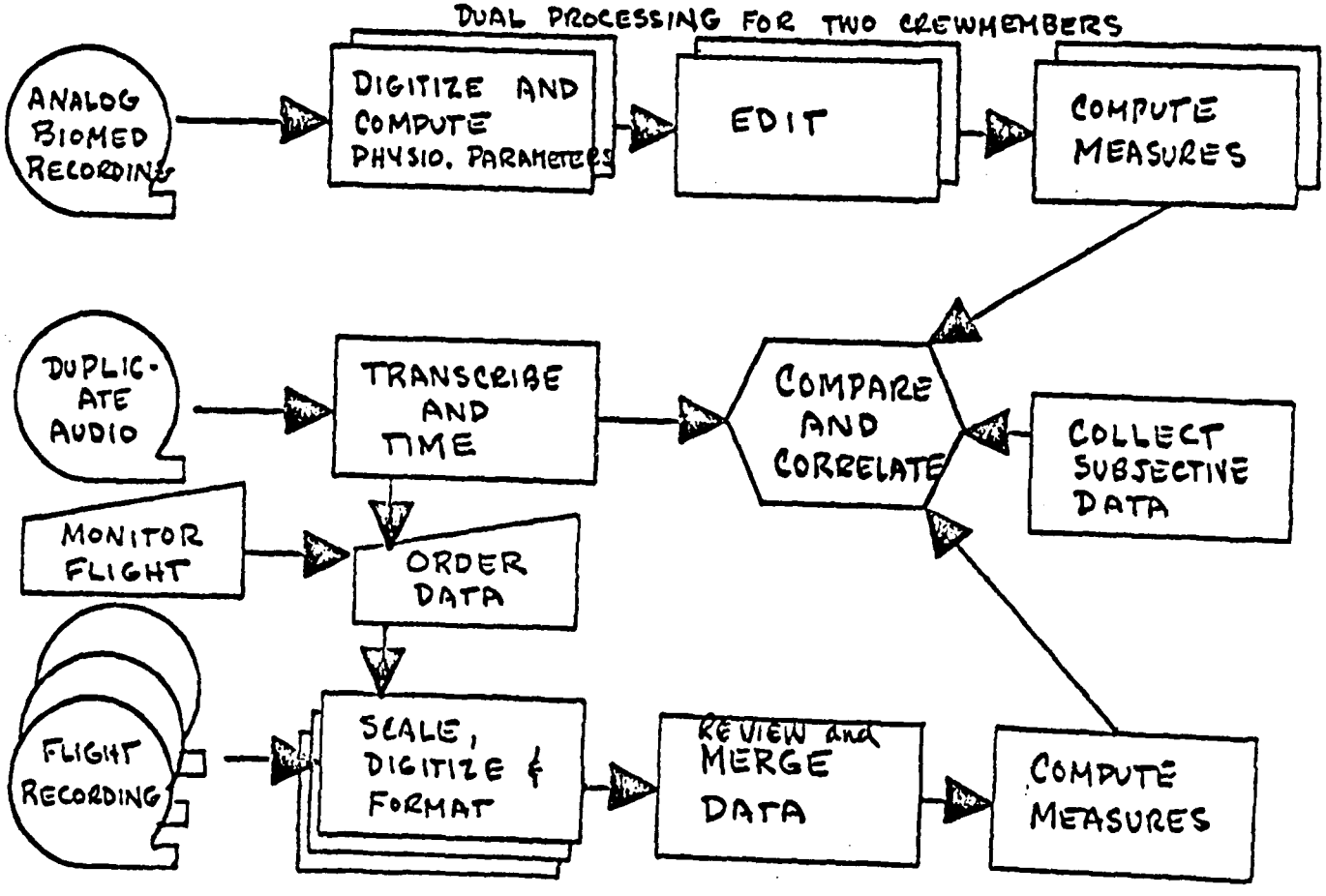


Figure 2. Example PMS for use in high-performance flight research.

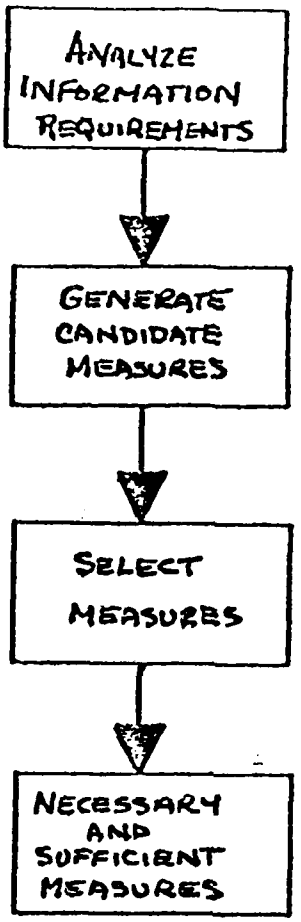


Figure 3. Selection of necessary and sufficient measures.

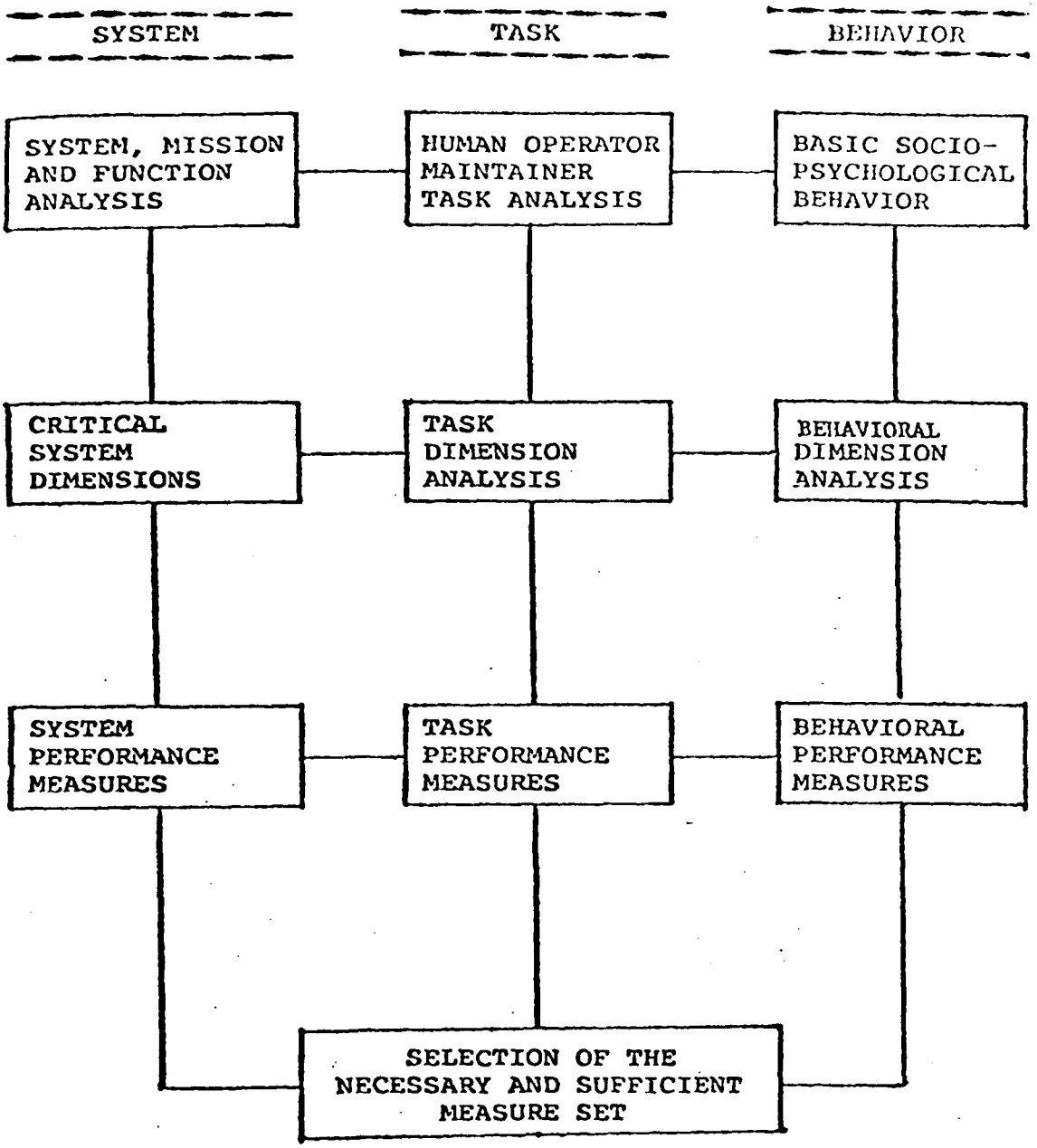


Figure 4. A general methodological model for the development of performance measurement.

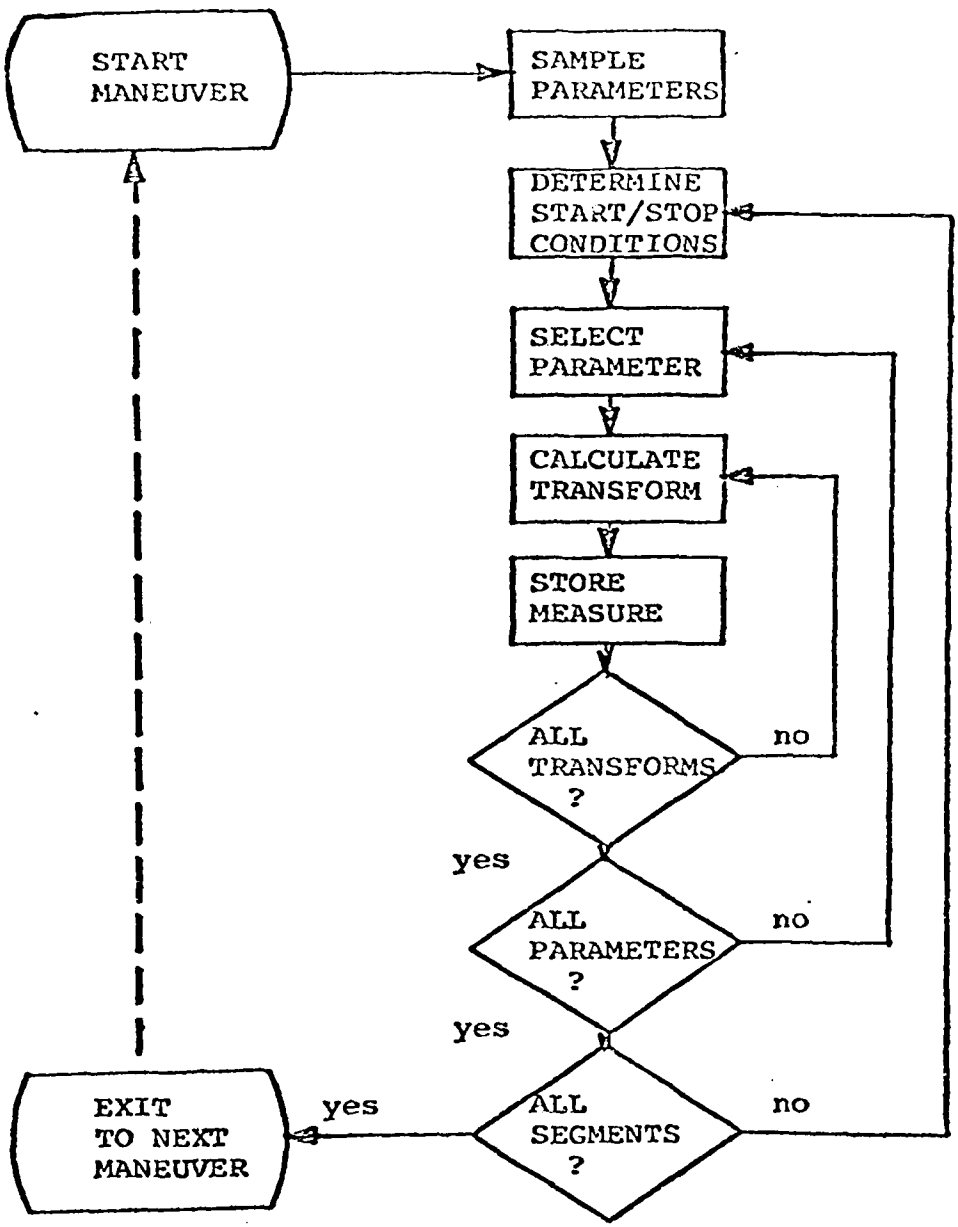


Figure 5. Block diagram for automated performance measurement computations.

REFERENCES

Blackman, R. B. and Tukey, J. W. The measurement of power spectra. New York: Dover, 1958.

Cronbach, L. J. and Gleser, G. C. Psychological tests and personnel decisions. Urbana, Ill.: University of Illinois Press, 1957.

Finley, D. L., Obermayer, R. W., Bertone, C. M., Meister, D. and Muckler, F. A. Human performance prediction in man-machine systems. The Bunker-Ramo Corporation, Canoga Park, Calif. National Aeronautics and Space Administration Contract NAS2-5038, August 1969.

Girshich, M. A. An elementary survey of statistical decision theory. Rev. Educ. Res., 1954, 24, 448-466.

Knoop, P. A. Development and evaluation of a digital computer program for automatic human performance monitoring in flight simulation training. Wright-Patterson Air Force Base, Ohio; Aerospace Medical Research Laboratories, AMRL-TR-68-971, August 1968.

Obermayer, R. W. Simulation, models, and games--Sources of measurement. Human Factors, 1964, 6(6), 607.

Obermayer, R. W. and Nicklas, D. Pilot performance measurement study. Final Technical Report, National Aeronautics and Space Administration, Contract NAS4-1919, Manned Systems Sciences, Northridge, Calif., June 1973.

Vreuls, D., Obermayer, R. W., Goldstein, I. and Lauber, J. W. Measurement of trainee performance in a captive rotary-wing device. NAVTRAEQUIPCEN 71-C-0194-1 U. S. Naval Training Equipment Center, Orlando, Florida, July 1973.

Vreuls, D. and Obermayer, R. W. Trainee performance measurement development using multivariate measure selection techniques. Manned Systems Sciences, Inc., Northridge, Calif. NAVTRAEQUIPCEN Report 73-C-0066-1, April 1974.

Wald, A. Statistical decision functions. New York: Wiley, 1950.

ABOUT THE AUTHOR

Mr. Richard Obermayer is Head, Laboratory Support Office, Navy Personnel Research and Development Center, and is currently responsible for research and development on methods for error reduction in computer data entry systems. He has over 23 years experience in human engineering and man-machine systems, with a long-term emphasis on human performance measurement He received BSEE and MSEE degrees from the University of Illinois in 1956; he is a member of the Institute of Electrical and Electronic Engineers and the Human Factors Society.

AUTOMATION OF PERFORMANCE MEASUREMENT

Robert C. Williges
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

ABSTRACT

Detailed prediction schemes of operator performance do not exist, even though there is a widespread requirement for quantitative performance measures throughout the life cycle of personnel systems. Some approaches to quantitative performance measurement are reviewed, and a description of a prototype performance assessment system based on polynomial regression prediction equations is presented. These equations can be used to determine tradeoffs in system design, to optimize personnel performance through the appropriate design of tasks, to isolate potential training requirements, and to provide a comparison standard for assessing personnel readiness in operational systems. In addition, a discussion is provided on the use of an automated performance measurement schema to enhance personnel effectiveness by embedded performance measurement, evolutionary system operation, and the development of appropriate job aids. Several unresolved issues dealing with future analytical and research needs are presented. These issues include the development of a prototype assessment system, an investigation of efficient strategies for data collection, the consideration of the appropriate candidate systems for automated performance measurement, and the cost effectiveness of such measurement systems. It was concluded that current advances in behavioral research methodology now make it feasible to consider the development of a complex, automated performance measurement system. Once this system exists, it can be used to generate a realistic data base of complex human performance from which meaningful generalizations can be made.

INTRODUCTION

The keystone for meaningful human factors applications to Navy personnel systems is the development and use of valid and reliable quantitative measures of operator performance. These measures are required throughout the life cycle of personnel systems. For example, predictions of personnel performance capabilities and limitations are used during systems design to maximize operator/machine interface, measures of personnel effectiveness are critical to operational test and evaluation of new systems, and continual assessment is needed of the operational readiness of Navy personnel in existing systems.

Given the importance of the requirement for evaluation of personnel performance, it is somewhat surprising that detailed, quantitative prediction schemes of the operator do not exist. Human factors handbooks (e.g., Parker and West, 1973; VanCott and Kinkade, 1972) and textbooks (e.g., McCormick, 1976) provide a great deal of information about human performance, but this information is not usually presented in a format that represents the complexity of the operational system.

PRECEDING PAGE BLANK-NOT FILMED

Recently, Blanchard (1975) conducted a survey of the Navy's research and development community to develop guidelines for a human resources data storage system. Part of this survey was directed toward the perceived utility of currently available human performance data sources. These sources included the experimental literature, human factors guides and manuals, and fleet exercise data. All three sources were found to be quite lacking. Most respondents reported that the experimental literature was, for the most part, not generalizable to applied work on Navy systems, whereas human engineering guides and manuals were only of limited utility. Fleet exercise data, on the other hand, were judged to be somewhat undependable because they were obtained from human observers and were subject to human errors and biases. In summary, Blanchard (1975) reported that many users of operational personnel performance suggested that objective, instrumented data collection regimes are needed.

The major reason for the present dearth of quantitative information is methodological. Personnel performance is a complex function of several parameters interacting simultaneously to manifest a particular behavior. The effect of one parameter, such as sonar display format, may change dramatically as a function of other variables, such as time-on-watch, background illumination, data update rates, etc. Consequently, a systems designer must have knowledge of the complex interrelationships of all of the critical variables in a system before totally accurate extrapolations to operator performance can be made. Traditional factorial design research methods cannot be used to gather these data due to the tremendous amount of time and cost necessary to evaluate all of the possible interactions that exist.

APPROACHES TO QUANTITATIVE PERFORMANCE MEASUREMENT

Several approaches have been used with limited success in providing quantitative measures of performance assessment. The more successful procedures have attempted to incorporate the complexity and multivariable aspects of personnel performance. The most recent approaches take advantage of computer capabilities to automate the performance measurement procedures.

Synthetic Tasks

Alluisi (1967) and his associates used a multiple-task performance battery in a synthetic work environment to assess complex performance. The tasks used in their battery were designed to measure functions typically demanded of the human operator. They included watchkeeping of static and dynamic processes; sensory-perceptual functions of signal discrimination and identification, long- and short-term memory functions; information reception and transmission functions; intellectual functions of information processing, decision making, and problem solving; perceptual-motor functions; and procedural functions. These functions were combined in synthetic work environments that consisted of passive and active tasks occurring in realistic workload arrangements. Subsequent measures of operator performance in these synthetic tasks could then be used to assess the effect of various critical functions that the human operator must perform in a variety of human operator/machine situations.

Factor-Analytic Approach

Another approach has been directed toward determining an empirically derived task taxonomy that can be used to describe complex tasks. This approach used factor

analytic procedures to specify the task dimensions empirically, as opposed to the Alluisi (1967) approach of first specifying general categorical terms of tasks. Fleishman (1967) reviewed his approach, which has been used in many laboratory and field settings to investigate skill learning, individual differences, operational task proficiencies, and the effect of changing task requirements. He was able to isolate various psychomotor factors and physical proficiency factors, which account for a wide range of perceptual-motor task performances. Some of these factors are summarized in the following lists (Fleishman, 1967):

Psychomotor Factors	Physical Proficiency Factors
1. Control Precision	1. Extent Flexibility
2. Multilimb Coordination	2. Dynamic Flexibility
3. Response Orientation	3. Explosive Strength
4. Reaction Time	4. Static Strength
5. Speed of Arm Movement	5. Dynamic Strength
6. Rate Control	6. Trunk Strength
7. Manual Dexterity	7. Gross Body Coordination
8. Finger Dexterity	8. Gross Body Equilibrium
9. Arm-Head Steadiness	9. Stamina
10. Wrist, Finger Speed	
11. Aiming	

Automated Performance Measurement

The two previous approaches developed a performance measurement battery based on either a conceptualized set of human operator performance functions (Alluisi, 1967) or an empirically derived set of underlying task factors (Fleishman, 1967) Although both of these approaches have been somewhat successful in explaining the characteristics of operator performance, neither has resulted in widespread operational system application. A third alternative, which is more task specific, automatically measures a variety of task parameters present in the actual system and then relates these parameters to human operator performance. Knoop (1973) developed such an automated performance measurement technique for assessing pilot proficiency in aircraft simulator and flight environments. To demonstrate the feasibility of this approach, she automatically recorded 20 flight variables (e.g., airspeed, pitch, RPM, throttle positions, flap positions) during Lazy 8 and barrell roll maneuvers in an instrumented T-37B aircraft. A variety of computer-aided techniques using Boolean functions and a computer analysis of maneuvers based on a priori hypotheses of potential measures were used to assess pilot performance. Although the performance measurement scheme was successful, the tremendous amount of complexity in defining the computer analysis may limit the widespread application of this approach.

An alternative procedure was used by Leshowitz (1976). He used a multiple regression approach to relate various flight parameters, which were automatically recorded in a low fidelity flight simulation, to instructor pilot ratings of several standard instrument maneuvers. He was able to predict the instructor ratings quite accurately using these regression techniques. Once equations such as these are validated, they can then be programmed easily into the simulator computer to provide automatic performance assessment.

Portable systems using minicomputer and microprocessor technology have been developed to collect and to analyze operator performance. One such system is described by Urmston (1975) for use with tactical display systems on Navy ships. The Operational Performance Recording and Evaluations Data System (OPREDS) is self-contained and does not interfere with normal tactical operations. The data collection system can automatically record console actions as well as instructions from the tactical displays system computer. In addition, a data reduction system can analyze the data to reconstruct the operational exercise into a complete set of action records. These records, in turn, can be used to evaluate operator performance. Additional work, however, is needed to develop measures of optimal performance as well as the relative influences of the various parameters in the action records on operator performance.

AUTOMATED PERFORMANCE ASSESSMENT SCHEMA

Recent advances in behavioral methodology now make it feasible to consider the generation of complex data bases of human performance. With the advent of high-speed, low-cost minicomputers and microprocessors, it is also possible to consider computer-based procedures for assessing and predicting personnel effectiveness. These capabilities, coupled with new research approaches, can be used to generate meaningful, quantitative prediction equations of operator performance. It is hoped that, once a variety of tasks is characterized in this manner, a realistic human performance data base will exist so that meaningful generalizations can be made to the design of new systems.

Performance Prediction Equations

In addition to measuring operator performance objectively, an automated performance assessment scheme must incorporate procedures whereby the effect of a system variable on resulting human operator performance can be expressed quantitatively. These mathematical relationships can then be used to predict human performance, to determine the set of systems parameters that results in optimum performance, and to provide performance standards for comparison purposes.

Polynomial Regression Equations. Traditional analysis of variance designs do not provide the appropriate data analysis procedures for expressing these quantitative functional analyses. Finkelman, Wolf, and Friend (1977), however, point out that polynomial regression procedures provide a reasonable alternative to analysis of variance for data characterized by lower-order trends. A polynomial expression provides a convenient approximation to a variety of mathematical relationships, thereby making it a powerful tool for predicting operator performance while still using a single standard format.

Polynomial regression prediction equations are extensions of general multiple linear regressions of the form

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon \qquad (1)$$

where Y is the measure of personnel performance (e.g., number of errors, efficiency rating, time to complete a task); X is the set of quantitative parameters (e.g., system response time, display size, operator duty time) that constitute the operational task; and βs are the partial regression weights for each term in the prediction equation. These partial regression weights can be determined by empirically using standard least squares procedures.

Often, system parameters interact and have curvilinear relationships with operator performance. Polynomial expressions can easily be expanded to include these higher-order curvilinear effects in addition to the first-order linear effects. Consider a complete second-order polynomial expression including three X variables: Such an expression might be used, for example, to assess the human operator's target detection latency (Y) on a visually time-compressed radar display as a function of target velocity (X_1), number of stored frames (X_2), and amount of clutter (X_3). The resulting second-order polynomial expression would be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1^2 + \beta_5 X_2^2 + \beta_6 X_3^2 \qquad (2)$$

$$+ \beta_7 X_1 X_2 + \beta_8 X_1 X_3 + \beta_9 X_2 X_3$$

where the β_1, β_2, and β_3 terms express the linear main effects of the three variables; β_4, β_5, and β_6 terms express the pure quadratic main effects; and β_7, β_8, and β_9 represent the linear X components of two-way interactions. Additionally, this polynomial expression can be extended to fit higher-order effects (e.g., cubic or quartic) of main effects and interactions merely by adding more terms to the expression.

Data Collection Procedures

Before a least squares analysis can be conducted to determine the partial regression weights given in Equations 1 and 2, various measures of the Y and X values must be gathered. Most regression analyses are conducted on data gathered in a passive manner. In other words, measures of the dependent variable, Y, are obtained simultaneously with the recording of the various values of the system variables, X. Across subjects and/or repeated observations there are enough fluctuations in the values of the X variables to conduct the regression analysis. The data recording scheme used by Leshowitz (1976) to predict instructor pilot ratings is an example of this approach. He used no systematic, experimental manipulations of the X variables to form the data base for the regression analysis.

An attractive alternative is to employ experimental designs to manipulate the various levels of the X variables. In this way, economical data collection schemes can be used to sample the effective range of interest of each X variable so that the optimal points of Y performance can be determined efficiently. For example, Gallaher, Hunt, and Williges (1977) used such an experimental manipulation approach with a fractional-factorial design to calculate polynomial regression equations for generating predictor display symbology.

Response Surface Methodology. One of the more perplexing methodological problems in using experimental manipulations to provide the data for polynomial expressions is how to collect the data in an efficient manner on operational tasks characterized by many interacting variables. The use of complete factorial multivariable analysis of variance design as a data collection scheme is impractical in these situations because of the large number of resulting treatment conditions. One particularly promising solution to this methodological problem is response surface methodology (RSM), which was originally introduced by Box and Wilson (1951).

These procedures provide an economical strategy employing a variety of techniques for conducting multi-factor research to seek an optimal level of performance. The strategy includes designs for initial pretesting, for collecting and analyzing data in stages to add and subtract variables of importance, and for determining the point or points of optimal performance.

Central-Composite Designs. One of the more useful designs in RSM is the central-composite design that was developed by Box and Wilson (1951). (Details on the description of this design as well as numerical procedures of RSM are documented in several sources, such as Box and Hunter, 1957; Clark and Williges, 1973; Cochran and Cox, 1957; Davies, 1954; Myers, 1971; Williges, 1976; and Williges and Simon, 1971). Central-composite designs require fewer data points than do comparable factorial designs, thereby allowing for the inclusion of more factors in the design. The design itself is a composite of two-level factorial or fractional-factorial designs augmented by a center point and 2K additional (star) points. Each of the K factors in the design occurs at five levels as shown in the three-factor, central-composite design depicted in Figure 1. With the appropriate replications, this design would provide the investigator with enough data to fit the complete second-order polynomial expression shown in Equation 2 and still be able to test for the possibility of higher-order effects. Not only does this design allow for collecting the data in stages using blocking procedures, but it also is possible to conduct subsequent analysis of variance tests on the resulting polynomial expression.

Although these design procedures have been used quite successfully in the chemical industry for several years, they have only recently been used in human factors research. During the last 6 years, several modifications of central-composite designs have been proposed to make them amenable to behavioral research, and these procedures have been used for predicting human performance in complex system environments. For instance, central-composite designs were used to predict human operator performance in the assessment of television-projected cartographic image displays (Williges and North, 1973) and in the evaluation of visual time-compression of complex radar displays (Clark, 1976; Mills and Williges, 1973; and Scanlan, 1975). Figure 2 provides examples of representative first- and second-order polynomial regression equations resulting from these applications. Consequently, it appears that applications of this procedure could provide viable procedures for generating complex prediction equations of personnel performance.

Personnel Assessment Schema

An automated performance assessment schema could easily consist of a set of polynomial regression equations of the type specified in Equation 2, which relates personnel performance to quantitative task parameters. Experimental manipulations

of these task parameters would provide the data necessary to generate the polynomial regression equations. These resulting equations, in turn, would be an important tool in determining optimal performance for use in design trade-off studies. In addition, such an approach would provide a realistic data base of operator performance in which many potentially interacting human operator, task, and environmental variables are considered simultaneously. By employing this same regression analysis across a variety of tasks, direct comparisons can be made of the relative importance of system parameters in a variety of tasks. These data could then be generalized to specify potential variables of interest across various categories of tasks and to project prediction equations in the design of new systems.

Automated performance assessment using polynomial regression techniques could be used for a variety of personnel tasks in the Navy inventory, but they are particularly useful in human operator/computer interface tasks where data pick-offs of various system parameters and operator performance measures are readily available. Several types of human operator/computer interface tasks are beginning to play a major role in the present and future operational tasks of the Navy. For instance, computer transaction tasks (including personnel data records, work scheduling, and equipment inventories) are quite prevalent. Real-time computer updating tasks in combat information centers using tactical information displays are primarily computer-based. Additionally, computer-controlled systems are present in most tactical and navigational displays in ship and aircraft systems. All of these tasks appear to be potential candidates for automated performance assessment procedures.

APPLICATIONS OF AUTOMATED PERFORMANCE ASSESSMENT METHODS

Experimentally derived regression equations representing the interrelationships of several quantitative variables affecting human operator performance can be used in several ways: Once the equations are developed, they can be used to determine tradeoffs in system design, to optimize personnel performance through the appropriate design of the task, to isolate potential training requirements, and to provide a comparison standard for assessing personnel readiness in operational systems. In addition to these important applications, automated performance assessment procedures offer some unique applications to human operator machine systems to enhance personnel effectiveness.

Embedded Performance Measurement

One unique feature of automated performance assessment procedures such as the polynomial regression approach is that they can be completely embedded in the operational task, particularly when the task is computer interfaced. The operational computer can be used for automatic data collection of both personnel performance and systems parameters in a fashion analogous to the concept of embedded training (Germas, Johnson, and Baker, 1976) in which the operational computer is used to train system personnel on the system operation during off-peak use periods. Alternatively, minicomputers and microprocessors can be used as relatively inexpensive, stand-alone data recording devices similar to the approach described by Urmston (1975). Obviously, the exact set of variables and the relationship among the variables in the embedded performance assessment package must be determined before economical measurement routines are embedded into the operational system.

Evolutionary System Operation

Box (1957) outlined a procedure called evolutionary operations, in which chemical process improvements are explored during the normal course of production by plant personnel. Various built-in procedures are used to increase productivity without interfering with the normal operations of the plant. Essentially, this procedure introduces a series of small, controlled variations into the normal operating cycle. The effects of these small (but systematic) changes are summarized in tabular form using simple statistical methods. This tabular information can then be used by the production manager as an aid in deciding whether to modify the operation or to wait for additional information. If he decides to modify production, he can choose one of the variants as the new production procedure and then restart the evolutionary process, change the pattern of variants in an indicated favorable direction, or choose new variables for manipulation. In this way, improvements in the production system are being explored continually without interference to the normal production routine.

Given the existence of an automated performance measurement system in which quantitative measures of personnel performance are available, an analogous evolutionary operations approach could be used to increase productivity in human operator/machine systems. The performance measurement system could be used to define the data base for choosing variables to be manipulated. In addition, much of the evolutionary operation could be automated using the performance measurement routine. In this way, the operational system would be constantly evolving into a more productive system without interfering with normal operations.

Performance Enhancement Procedures

Other human operator performance enhancement procedures can be considered using online performance monitoring procedures. For example, Enstrom and Rouse (1976) developed an online fading-memory system identification model using linear discriminant analysis to determine how the human operator allocates attention between control and monitoring tasks. This procedure can be used to allocate decision-making responsibilities in human/computer tasks. In addition, Poulton (1973), in a review of fatigue in vigilance research, describes a procedure whereby inspector behavior can be improved through the use of augmented feedback. Inspector performance is continually monitored and the appropriate feedback is provided contingent upon the actual performance levels. Clearly, automated performance measurement is a prerequisite to instrumenting these augmented feedback systems.

Obermayer (1977) described the implications of automating a data entry subsystem to reduce the errors and labor-intensive effort encountered in large-scale Navy personnel information systems. Data entry errors are particularly prevalent on the optical character recognition (OCR) forms used in these systems. Through the use of automated performance measures, both design criteria and candidate system variables responsible for data entry errors in OCR forms can be isolated. This information can then be used to help formulate automated job aids and work simplification procedures to improve operator performance.

SOME UNRESOLVED ISSUES

Even though the use of automated performance measures offers the potential for significant improvement in system operations, and polynomial regression seems to be a logical candidate for developing a measurement system based on functional relationships, these performance assessment systems do not exist. Not only do the necessary data bases need to be developed, but several pertinent analytical and research issues need to be considered before any large-scale automated performance measurement schema is developed.

Prototype Assessment System

Before implementing a large-scale, complex performance measurement system, a prototype system needs to be developed to demonstrate the feasibility of using a polynomial regression approach. Both the reliability and the validity of this approach can be evaluated in the prototype system. During the development of this system, some of the limitations of regression procedures can be evaluated. For example, polynomial regression is most useful in considering continuous, quantitative variables as opposed to discontinuous or qualitative variables. Even though separate regressions can be derived for qualitative variables as Williges and North (1973) demonstrated in considering color versus black/white television systems (see two such examples in Table 2), the regression approach becomes unwieldy when many nonquantitative variables are critical to system operation. Consequently, a prototype system needs to be developed as a design guide for future development of operational performance measurement systems.

Strategy for Data Collection

Several alternative procedures exist for collecting the data required for solving the prediction equations. The central-composite designs represent only one of these approaches. More important, the entire strategy for economical data collection in complex human operator/machine environments needs to be considered. Simon (1973), for example, recommended a three-stage process for screening a large number of factors. The first stage includes saturation designs composed primarily of two-level, fractional-factorial designs. Augmentation is used in the second stage to separate main effects from the two-factor interactions. And, the final stage includes additional data that isolate two-factor interactions that are important.

In addition to the overall strategy for experimentation, corollary issues need to be considered. Most research strategies address themselves to procedures for adding and subtracting independent variables rather than dependent variables. Ways of choosing and combining several measures of operator performance also need to be explored. Multivariate procedures such as canonical analysis (Morrison, 1967) may provide a potential way of assessing the combined effects of several dependent variables.

Types of Personnel Systems

Obviously, a polynomial regression approach will not be a panacea for all automated performance assessment situations, but it should be applicable to a large variety of personnel tasks. Analytical studies are needed to determine which performance assessment procedures are the most appropriate for various classes

of personnel systems. Human operator/computer interface tasks appear to be a prime candidate for regression-type assessment protocols, but even these tasks need closer consideration. For instance, a regression procedure may be more or less effective in event-based computer tasks, such as a personnel data entry system, as opposed to a time-based computer task, such as sonar display monitoring. The real-time aspects of many human operator/computer tasks place additional constraints on the personnel assessment requirements in terms of data rates and variables to sample.

Cost Effectiveness

The benefits in terms of cost effectiveness must be evaluated in considering any type of automated performance measurement system. The costs for developing a comprehensive automated assessment system are substantial, but the potential benefits in terms of enhanced system design and improved operational performance should more than outweigh these development costs. Due to the extremely large expense of operating any labor-intensive system, an operational performance system that resulted in even a small improvement would be cost effective.

CONCLUSION

Automated performance measurement schemes need to be developed for personnel systems. Until these systems are developed, there seems to be little hope of generating and using realistic human performance data bases that characterize the complex multivariable interrelationships of real-world systems. Given the recent advances in behavioral research methodology, it now seems feasible to consider automated performance assessment. A polynomial regression prediction equation applied to human operator/computer interface tasks appears to be the most promising candidate for developing the prototype performance measurement system. Once such a system exists, it can be used in a variety of ways, including embedded performance measurement, evolutionary system operation, and enhanced performance procedures to increase productivity in personnel systems. In addition, it will provide the necessary data base from which theoretical extrapolations can be made to the design of future systems.

ACKNOWLEDGMENTS

Contractural support for this paper was provided by the Navy Personnel Research and Development Center under Contract Number N00123-77-C-1081. Dr. Frederick A. Muckler served as the scientific monitor for this contract.

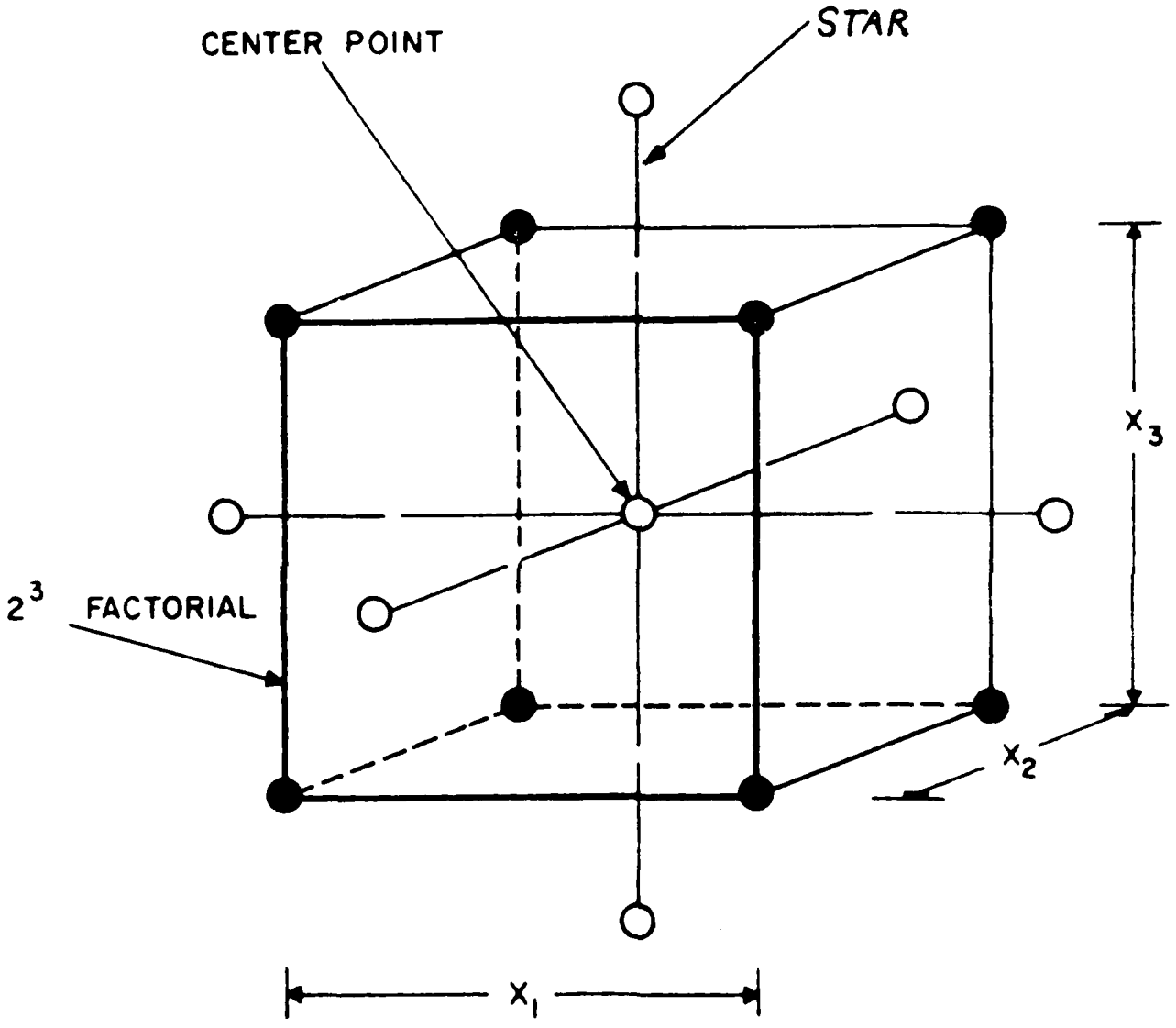


Figure 1. Unique data points of a three-level central composite design.

Video Cartographic Symbol Location Performance (Williges and North, 1973)

$$CL\ (\text{Black and White Monitor}) = 1.69 + 0.19F + 0.33D + 0.06V + 0.11T$$

$$CL\ (\text{Color Monitor}) = 1.62 + 0.36F + 0.19D + 0.17V + 0.22T$$

where CL = correct location
F = camera focus
D = density of nontarget symbols
V = visual angle
T = TV raster lines/mm of map

Radar Target Detection in a Simulated Surveillance System (Mills and Williges, 1973)

$$P(CI) = .293 + .2193BSR - .023TIR - .303CRP - .002CD + .0009TV$$
$$- 1.285BSR^2 - .128BSR \times TIR + .290BSR \times CRP + .0002\ BSR \times CD$$
$$+ .0002BSR \times TV - .004TIR^2 + .032TIR \times CRP - .0002TIR \times CD$$
$$+ .00003TIR \times TV - .090CRP^2 + .0002CRP \times CD - .00002\ CRP \times TV$$
$$+ .00001CD^2 - .0000004CD \times TV - .000001TV^2$$

where P(CI) = probability of correct track initiation
BSR = blip/scan ratio
TIR = target introduction rate
CRP = clutter replacement probability
CD = clutter density
TV = target velocity

Figure 2. Examples of polynomial regression prediction equations of human operator performance.

REFERENCES

Alluisi, E. A. Methodology in the use of synthetic tasks to assess complex performance. Human Factors, 1967, 9, 375-384.

Blanchard, R. E. Human performance and personnel resource data store design guidelines. Human Factors, 1975, 17, 25-34.

Box, G. E. P. Evolutionary operation. A method for increasing industrial productivity. Applied Statistics, 1957, 6, 81-101.

Box, G. E. P. and Hunter, J. S. Multifactor experimental designs for exploring response surfaces. Annals of Mathematical Statistics, 1957, 28, 195-241.

Box, G. E. P. and Wilson, K. B. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society, Series B. (Methodological), 1951, 13, 1-45.

Clark, C. D. Mixed-factors central-composite designs: A theoretical and empirical comparison. Savoy, IL: University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-76-13/AFOSR-76-6, August 1976.

Clark, C. and Williges, R. C. Response surface methodology central-composite design modifications in human performance research. Human Factors, 1973, 15, 295-310.

Cochran, W. G. and Cox, G. M. Experimental designs. (2nd ed.) Chapter 6A. Factorial experiments in fractional replications. New York: Wiley, 1957, 244-292.

Davies, O. L. (Ed.) The design and analysis of industrial experiments. Chapter 11. The determination of optimum conditions. London: Oliver and Boyd, 1954, 495-578.

Enstrom, K. D. and Rouse, W. B. Telling a computer how a human has allocated his attention between control and monitoring tasks. Proceedings of Twelfth Annual Conference on Manual Control. Moffett Field, CA: Ames Research Center, NASA TMX-73, 170, May 1976.

Finkelman, J. M., Wolf, E. H., and Friend, M. A. Polynomial regression analysis as an alternative to ANOVA for data characterized by lower-order trends. Human Factors, 1977, 19, 279-281.

Fleishman, E. A. Performance assessment based on an empirically derived task taxonomy. Human Factors, 1967, 9, 349-366.

Gallaher, P. D., Hunt, R. A., and Williges, R. C. A regression approach to generate aircraft predictor information. Human Factors, 1977, 19, in press.

Germas, J. E., Johnson, E. M., and Baker, J. D. Embedded training: Using a computer system to train the system user. Paper presented at 6th Congress of the International Ergonomics Association, July 11-16, 1976, College Park, MD.

Knoop, P. A. Advanced instructional provisions and automated performance measurement. Human Factors, 1973, 15, 583-597.

Leshowitz, B. Personal communication regarding AFOSR-73-2555 contract, Adaptive procedure for measurement of flying training proficiency, 1976.

McCormick, E. J. Human factors in engineering and design. New York: McGraw-Hill, 1976.

Mills, R. G. and Williges, R. C. Performance prediction in a single-operator simulated surveillance system. Human Factors, 1973, 15, 337-348.

Morrison, D. F. Multivariate statistical methods. New York: McGraw-Hill, 1967.

Myers, R. H. Response surface methodology. Boston: Allyn and Bacon, 1971.

Obermayer, R. W. Accuracy and timeliness in large-scale data-entry subsystems. Proceedings of Twenty-first Annual Meeting of the Human Factors Society, Santa Monica, CA: Human Factors Society, October 1977.

Parker, J. F. and West, V. R. (Eds.) Bioastronautics data book. Washington, DC: U. S. Government Printing Office, 1973.

Poulton, E. C. The effect of fatigue upon inspection work. Applied Ergonomics, 1973, 4, 73-83.

Scanlan, L. A. Apparent motion quality and target detection on a visually time-compressed display. Savoy, IL: University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-75-16/AFOSR-75-6, November 1975.

Simon, C. W. Economical multifactor designs for human factors engineering experiments. Culver City, CA: Hughes Aircraft Company, Technical Report Number P73-326A, June 1973.

Urmston, R. Operational performance recording and evaluation data systems (OPREDS) (descriptive brochure) San Diego: Naval Electronics Laboratory Center, 1975.

Van Cott, H. P. and Kinkade, R. G. (Eds.) Human engineering guide to equipment design. Washington, DC: U. S. Government Printing Office, 1972.

Williges, R. C. Research note: Modified orthogonal central-composite designs. Human Factors, 1976, 18, 95-98.

Williges, R. C. and North, R. A. Prediction and cross-validation of video cartographic symbol location performance. Human Factors, 1973, 15, 321-336.

Williges, R. C. and Simon, C. W. Applying response surface methodology to problems of target acquisition. Human Factors, 1971, 13, 511-519.

ABOUT THE AUTHOR

Robert C. Williges is a professor of Industrial Engineering and Operations Research as well as a professor of Psychology at Virginia Polytechnic Institute and State University. He is a fellow of both the American Psychological Association and the Human Factors Society. Currently, he is serving as Editor of Human Factors. In 1973 he won the Society's Jerome H. Ely award for the best paper published in Human Factors. From 1968 to 1976 he was on the faculty at the University of Illinois at Urbana-Champaign where he held academic appointments in psychology and aviation. In addition, he served as the associate head for research of the Aviation Research Laboratory and assistant director of the Highway Traffic Safety Center at the University of Illinois. He received his A.B. degree in psychology from Wittenberg University in 1964 and his M.A. and Ph.D. degrees in engineering psychology from The Ohio State University in 1966 and 1968, respectively. His research interests include team training, visual monitoring of complex computer-generated displays, inspector behavior, behavioral applications of central-composite designs, transfer of training, motion and visual cues in simulation, adaptive training procedures, and assessment of human performance in complex system operation including investigation of rate-field, frequency-separated, visually time-compressed, and predictor displays.

SELECTING PERFORMANCE MEASURES: "OBJECTIVE" VERSUS "SUBJECTIVE" MEASUREMENT

Frederick A. Muckler
Navy Personnel Research and Development Center
San Diego, California

ABSTRACT

The commonly held distinction between "objective" and "subjective" measurement is considered to be a pseudo-difference. All measurement is "subjective" in that human acts and judgments are involved in every step of the process of measurement; no measure exists outside the individual(s) who define, collect, and interpret data. Eight criteria may be applied for selecting performance measures: (1) validity, (2) reliability, (3) precision, (4) completeness, (5) generalizability, (6) non-reactivity, (7) utility, and (8) information needs. A decision process for applying these criteria to selecting performance measures is suggested.

OBJECTIVITY IN MEASUREMENT

"Objective" versus "Subjective" Measurement

In selecting appropriate measure sets for human performance measurement, several issues can be raised about the nature of the measures chosen. One all-pervasive issue that underlies all human measurement is the issue of "objective" versus "subjective" measurement. This is shown in the preference of some for measures that depict "what people do" rather than "what people say they do." But underlying this choice is the implicit assumption that "objective" measures are inherently superior to "subjective" measures. It may be worthwhile to question that assumption, or, at least to clarify what is meant by the distinction.

Modern science has presumably been built upon a fundamental premise of objectivity. That is, the theories and findings of science are presumed to be "true" and independent of the individuals who create the theories and collect the data. From this point of view the goal of science is universal truth uncontaminated by the biases of individual observers. To be "objective," then, is perhaps the greatest "good" in science. To be "subjective" is to introduce presumably biased and distorted estimates of "truth."

This basic philosophical assumption finds an immediate application in the question of appropriate measure sets. It is assumed that "objective" measures are "good" and "subjective" measures are bad. In classical psychometric theory it has been traditionally assumed that any obtained measure is in fact an obscure combination of a "true" score plus some random error of measurement on that score (cf., Guilford, 1954). Presumably, "objective" measurement is the only avenue to the "true" score while "subjective" measurement hopelessly confounds the obtained measure. Wile the error of measurement may not necessarily be random (cf., Cronbach, Gleser, Nanda and Rajaratnam, 1972) and may in fact result from many sources, the estimation of the "true" score can only come from "objective" measurement.

PRECEDING PAGE BLANK-NOT FILMED

In general, "objective" measurement is that obtained from measurement independent of the human observer. In so far as human actions and/or judgments are involved in the measurement process the measurement must be said to be "subjective." Based on that distinction it is the thesis of this paper that all measurement is "subjective" in all the sciences. Further, it is held that to assume that "objective" measurement in this sense is even possible is at best an unwarranted optimism and at worst a delusion. At the least, it is proposed that the distinction between "objective" and "subjective" measurement is not absolute but, at best, relative.

The Process of Measurement

To support these statements it is useful to look at the act or process of measurement in all sciences. Four sequential steps may be distinguished: (1) selecting measures, (2) collecting data, (3) analyzing data, and (4) interpreting data. It is assumed that measurement cannot occur without these steps.

Selecting Measures. In the more advanced of the sciences appropriate measures are usually dictated by quantitative theory. That is, the theory decides what parameters are measured and, by exclusion, what are not. Given the theory, measurement then becomes a purely mechanical process where the data become a test (confirming or denying) of the theory. This process Kuhn (1961) has termed "textbook measurement," and he takes some effort to show that actual measurement does not follow that simple model. Further, scientific theory is the product of some human mind(s), and, finally, this century has seen the constant change in theory in all scientific disciplines. "Objective" measurement, therefore, is dictated by a human theoretical framework that very probably will be changed resulting in new theory and measure sets. Only in the very long (historical) sense can it be hoped that this process will result in "true" scores.

For the most part, the behavioral and social sciences lack even the guiding hand of quantitative theory. Therefore, to measure at all, means taking some a priori assumptions for selecting a measure set. In performance measurement, that selection seems to be primarily dictated by convenience or selective interest. Data that are available become the core measure set or the investigator has some particular interest in some part of the problem for which he selected measures. This process of selection is surely judgmental and by no definition could be termed "objective."

Collecting Data. Standard texts of research methods place so much stress on error sources in data collection that one is assured that it is indeed a major problem. And much of the problem appears to be human error in what seems like a basically mechanical step. It might seem that in the quantitative sciences this problem is minimized by instrumentation but even if the measurement task is simply dial reading we may expect significant- and nonrandom-reading errors (McCormick, 1964). And, while corrected by modern instrumentation, it is perhaps instructive to recall the classic case in the eighteenth century of the errors in stellar transit recordings by the unfortunate Kinnebrook (Boring, 1950).

While there appears to be no data other than anecdotal, it would appear that to some extent human performance measurement may well be distorted intentionally in some situations. The uses to which human performance measurement are often put involve rather direct rewards and punishments for the individuals involved. It is perhaps not surprising that, unsupervised or checked, systematic distortion may appear in some kinds of data records on human performance.

Analyzing Data. One of the blessings of the current century is the birth and growth of inferential statistics. This is particularly true for the behavioral and social sciences where data in probabalistic form is the norm and, in particular for human performance measurement, where the data are always statistical in nature. But how data are analyzed and presented always involves the judgment of the individual researcher--as an examination of any current journal will confirm. The analytic techniques seem to be not so much a function of the analysis problem at hand but, rather, of the current familiarity of the researcher with the state of statistical method and/or the particular interest of the researcher in certain aspects of the data. Statistical texts would imply that analysis is a mechanical process; practice suggests that it is a judgmental process.

Interpreting Data. "What do the data mean?" Surely, of all the steps in the act of measurement human judgment is critical here. Interpretation is a human act--a subjective process. It is difficult to conceive of this step as "objective" in the sense defined before. With an analogy to psychometric methods (cf., Guilford, 1954, pp. 251-256), it is possible to define objectivity of interpretation based upon the degree of agreement between observers, and, if acceptable, that degree is measurable. But this is a social criterion and its validity in the sense of "truth" is subject to some question. What is the consensus today may well not be tomorrow.

Man as the Measurer. As Lorge (1967) and his associates have pointed out, the act of measurement means process, instrument, units, and results. The preceding discussion has attempted to demonstrate that the act of measurement inherently involves in every aspect human action and judgment, and, hence, the possibility of "subjective" elements. In the 5th Century B. C., Protagoras stated the famous dictum: "Man is the measure of all things." Klein (1974) would modify that statement to: "Man is the measurer of all things."

Human as Object and Measurer

One of the most difficult problems of human performance measurement is the inherent fact that the human is the object of measurement but, at the same time, the human may be used as the measuring device for that measurement. The question of "objective-subjective" measurement in this area appears to rest primarily on the validity of the object being used as a measuring device.

Consider the problem on productivity measurement. We ask the question: "How much has the individual produced in a given unit time?" Unless one accepts totally simplistic measures, this question turns out to be an extremely complex one. "Objective" measurement could concentrate on the quantity of products; what and how many were produced in the time span of interest? Even here without continuous measurement records are often difficult and costly to generate. Further, unless the individual is working completely alone, it is often difficult to partition the contribution the individual has made relative to those of others involved in the product.

For any product, the question of quality of that product must also be raised. Sheer quantity in production is an inadequate metric unless error rates are considered as well. For most activities in current industralized civilization the quality issue is very complex. For example, most products are now concerned with "services"; the question is not only how much but how good are the services?

From the measurement standpoint, the human observer is a potential source of data either for quantity or quality of productivity. The measurement question is: How adequate is the human as a measuring device? But even this question is not sufficient. Better, it is a question of which sets of measures, among those available and cost-effective, which can most effectively be used? For most judgments of quality of performance it would appear that some source of human measurement is currently preferable: either data from the individual, his/her peers, subordinates, and/or superiors. Parenthetically, this is not meant to necessarily imply a theoretical superiority; rather it is more a statement of current technology in performance measurement which is very much subject to change.

How adequate the human is as a measuring device has been the subject of much discussion (cf., Luce, 1972). The difficulties with the human as a measuring instrument are certainly well known. Chapanis (1959), for example, has commented on the problem that the human is not a good observer of complex events and this is probably particularly true with multi-dimensional phenomena occuring at relatively high rates. But a detailed technology and data base on the efficacy of the human as a measuring device is not available. Until (and if ever) such a technology is available, the human as a measuring device must be evaluated for each empirical situation on a judgmental basis. Whatever the case, the human is inherently a part of measurement, and his capabilities and limitations must be carefully assessed. In fact, it may well be in some specific cases that the human remains the measuring device of choice. That decision will rest on the application of several measurement criteria--the "rules" for selecting measure sets in specific applications. Some of those "rules" are discussed in the next section.

CRITERIA FOR SELECTING MEASURES

According to a famous statement by Stevens (1951), "Measurement is the assignment of numerals to objects or events according to rules." The question then is: What set of criteria or "rules" should guide measure selection in performance measurement? The previous discussion has attempted to show that the "objectivity-subjectivity" distinction is not a valid rule. In the following, some eight criteria are suggested, all of which should be considered in every application for personnel performance measurement. They are: (1) validity, (2) reliability, (3) precision, (4) completeness, (5) generalizability, (6) non-reactivity, (7) utility, and (8) information needs. These are not independent dimensions, and the selection of measure sets will depend upon the interactions between them.

1. Validity. It is traditional wisdom that the measures should measure what they say they measure. And much of the technology of modern psychometric theory (cf., Guilford, 1954) has been concerned with the development of a vast variety of sophisticated techniques to evaluate empirically the degree of validity a measure contains. Much of the older literature contained the implicit assumption that the higher the validity coefficient the better the measure, but Cronbach and Gleser (1965) have shown conclusively that that is not true when one considers the cost-effectiveness of measures and the uses to which the measures will be put. For most applications, very high validity coefficients (e.g., 0.70 and above) are neither practical nor desirable.

2. Reliability. For a measure to mean anything it must be repeatable or reproducible. This is a fundamental tenent of all scientific measurement (cf., Wilks, 1961). This assumes, however, that the process being measured remains

reasonably stable so that a measurement taken in time sequence will be of the same system and not one that has changed. In performance measurement and particularly with regard to productivity enhancement the system may indeed change and low reliability of measures may result. Two alternatives are possible: the measure may be unreliable under a constant system process or the measure may be unreliable because the process is different. Measure reliability during the human learning process is often low, not because the measures are poor but because the human system being measured is changing.

3. Precision. The term "precision" is sometimes used as synonomous with "reliability" (cf., Wilks, 1961, p. 6). As used here it refers to the level of accuracy required for a measure. With the general scientific emphasis on quantification it is often assumed that a very high level of numerical precision is required for measure sets. In scalar terms the goal often seems to be at least ratio scale measurement (cf., Stevens, 1951). But for many purposes this level of accuracy is not required. Consider, for example, performance appraisal where the need is simply to rank all employees in order of merit. In this case, an ordinal scale (assuming the ranks are valid and discriminable) is sufficient. Indeed, precision beyond that is unnecessary. The level of accuracy of a measure depends to a great deal upon the use to which that measure will be put.

4. Completeness. All behavioral and social sciences phenomena appear to be normally complex and multi-dimensional. Human performance measurement is definitely a case in point. The completeness criterion deals with the problem of the degree to which the measure set in fact measures the dimensions of the process. For example, if the goal is to measure the human, measurement of the finger no matter how valid, reliable, and precise the finger measure may be is not a complete measure set for the human. It would appear that most performance measurement sets tend to incompleteness. If so, it would appear that the principal reason is an incomplete knowledge of the phenomena or an incomplete understanding of the various levels of the process. Evans (1969), for example, has noted much confusion in the measurement of job satisfaction simply because different aspects of job satisfaction are not clearly separated.

Very frequently, in performance measurement, the basic process to be measured is not clearly defined dimensionally. Because this is so, Connelly (1974) and his associates have defined a computer-assisted performance measurement system where candidate measures can be established and measured and then either retained or discarded upon further investigation. It would appear in many cases of performance measurement that a cautious and adaptive approach to measurement might well be wisely used.

5. Generalizability. It is a widely held scientific goal that measurement in specific cases may be evaluated by the degree to which data may be generalized to other settings. While this may well be true with respect to basic knowledge about human performance derived from performance measurement, it may not be as important for specific applications. Indeed, one should be cautious in generalizing one set of measures from one situation to another.

Consider the problem of measuring and evaluating faculty effectiveness. Both the dimensions for evaluation and the weightings placed on those dimensions may vary widely from one academic situation to another. A measure set derived for a research institution may vary considerably when an institution which places prime importance on teaching is considered. Generalization from one system to another depends not so much on the measures as it does on the degree of similiarity between the systems.

6. Non-reactivity. One of the great modern notions of physics is the uncertainty principle: ". . . the energies necessary to observe the behavior of subatomic particles are so great that they distort or alter the objects observed" (Klein, 1974, p. 192). Psychologists have known for a long time that the act of measurement may interfere with the process being measured. One classic case, for example, has been in time-and-motion studies where the direct measurement of worker behavior can result in an immediate slowdown of that behavior. This is one of the primary reasons for the low reliability and validity of work standards generated by time and motion study (cf., Galvendy and Seymour, 1973, pp. 204-215). A correction for this problem has been the development of "unobtrusive measures" (Webb, Campbell, Schwartz and Sechrest, 1966) which attempt to measure behavior without the object of measurement being aware of it. While an extremely important development in measurement theory and practice, a very serious problem is now evident with these measures in light of the implications of the Privacy Act (cf., Sechrest, 1975).

It may not be, however, that non-reactivity of measurement is desirable. In the case of productivity enhancement, for example, the act of measurement of performance may be deliberately motivating. Further, using a program of management-by-objectives, the individual must know precisely on what dimensions measurement is being made and the standards of performance associated with each dimension. Here, measurement must necessarily "interfere" with the process being measured since the measurement is a part of the process.

7. Utility. No single subject in performance measurement generates more discussion and less detailed analysis than the question of utility of measurement. How much will the measurement cost? And, what benefits will be gained by performance measurement relative to the cost of measurement? That payoff function analysis of measurement can result in radically different and better measurement is well known (cf., Cronbach and Gleser, 1965). But rarely in practice does one see utility evaluation performed before a measurement set is put into practice. Not the least of the problems is to specify precisely what the value of measurement may be. And, for that matter, the calculation of costs turns out not to be simple, as well. But the value of utility analysis is very clear not only for practical reasons but for technical objectives as well. In every case, a utility analysis of a measure set will always result in radical revisions of the measure set selected.

8. Information Needs. The purpose of measurement is not to measure but to derive some kind of information. There are many purposes to which data may be put (cf., Guttman, 1971; Krantz, 1972). It seems imperative that the information needs be clearly specified as a part of the selection of any measure set. Yet, this analysis seems to be very rarely performed. It is not possible to select a rational measure without some kind of answer to the question: What do you want to know? It is not necessary that a definitive and specific answer be given to that question before measures are selected and used, but some level of answer must be found to dictate good measurement. Further, as Meister (1976) has pointed out, in the case of performance measurement, questions may be asked from many points of view: the system, the mission, and the individual. Productivity enhancement is usually presented from the system or mission point of view as a desirable objective of the organization, but it seems equally important to measure how the individual views the results of a productivity enhancement program.

Multiple-Measure Sets. Frequently, researchers and operational measurement specialists show a marked preference for one kind of measurement over another. This is most commonly heard, for example, in the preference for "objective"

measurement. At the present state of technology, however, a mixed measurement set may be the best possible. Sechrest (1975) has argued that non-obtrusive measures cannot be considered as alternatives or substitutes for questionnaires or interviews but as complementary measures ". . . which can strengthen our interpretations and lend greater confidence to our conclusions." In quite a different example, Meister (1978) has made a similar argument. For the generation of data banks for estimating human reliability, data generated from experimental studies, operational investigations, and expert opinion are all necessary to complete each other at the present state of knowledge. To the extent that a measure may enhance information needs, it is doubtful that any measure should be excluded--no matter how "subjective" that measure may appear.

USING THE CRITERIA IN MEASURE SELECTION

Figure 1 presents a suggested flow by which the eight criteria just discussed may be used in selecting measure sets for performance measurement. The figure shows basically four steps in selecting a measure set for a given system measurement problem. The most important, perhaps, is the initial definition of the measurement problem in terms of information needs and a specification of the dimensions to be measured. It is this part of the process that most frequently is ignored or is very obscure in practice. Simply put, the process initially asks two questions: (1) What do you want to know? and (2) What is it you want to measure?

A critically important feature of this recommended decision process is the specification of measurement dimensions. For each of the dimensions it is then possible to consider alternative measurement tools. As noted before, however, it may not be possible to define a specific set of dimensions. In that case, the technique of Connelly (1974) and his associates may be useful; indeed, it may be essential.

Once an alternative set of measuring tools has been defined it is then possible to apply six of the criteria: validity, reliability, precision, non-reactivity, generalizability, and utility. The nature of the evaluation is clearly that of a trade-off between these criteria. Unfortunately, no systematic method is known for such a trade-off although the work of Cronbach and Gleser (1965) is clearly of importance for validity, precision, and utility. Without a specifically defined procedure, the trade-off must be qualitative and judgmental. What would appear to be omitted in this set is concern with instrumentation. However, instrumentation is considered to be an inherent part of the utility analysis.

In any applied situation, performance measurement may be severely restricted by practical constraints of that situation. It may be necessary, for example, to minimize instrumentation and cost and actual data collectors may not be trained in the problems associated with collecting data. But the impact of these constraints will be felt directly on one or more of the six criteria used in evaluating alternative measure sets. It is to be expected that operational measurement is less than "ideal"; what is important is that one be able to know just what kind of measurement is taking place.

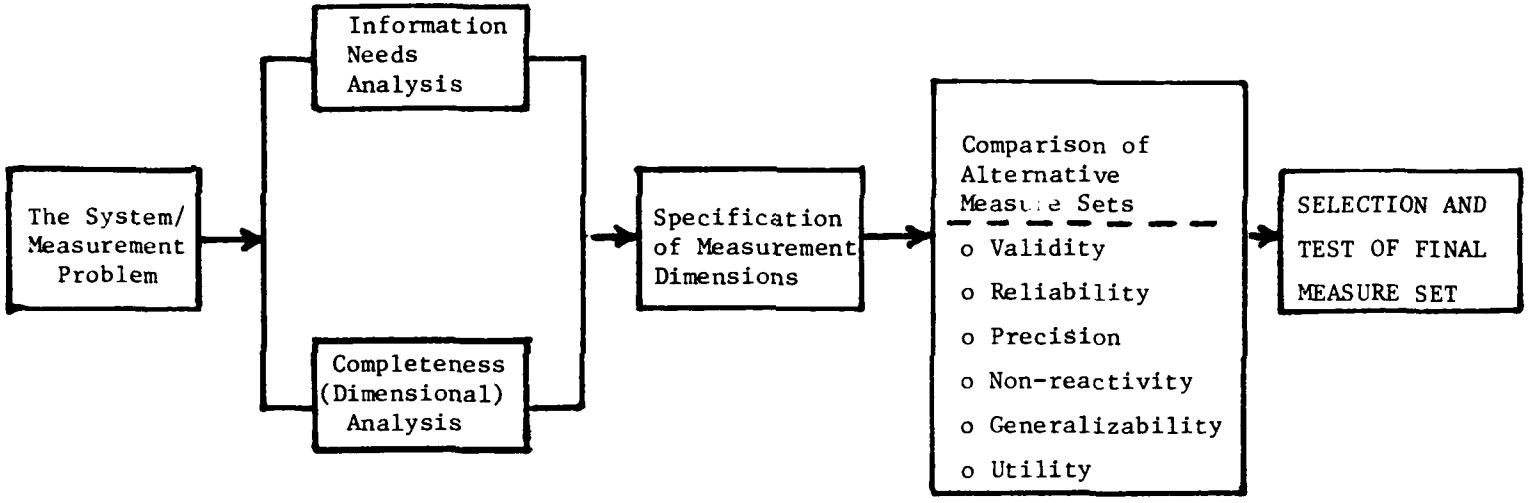


Figure 1. Suggested decision process for selecting measures in performance measurement.

REFERENCES

Boring, E. G. A history of experimental psychology. New York: Appleton-Century-Crofts, 1950. Chapter 8: The Personal Equation; Pages 134-153. (Second Edition)

Chapanis, A. Research techniques in human engineering. Baltimore: The Johns Hopkins Press, 1959.

Connelly, E. M., Bourne, F. J., Leontal, D. G. and Knoop, P. Computer-aided techniques for providing operator performance measures. United States Air Force, HRL TR 74-87, December 1974.

Cronbach, L. J. and Gleser, G. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

Cronbach, L. J., Gleser, G., Nanda, H., and Rajaratnam, N. The dependability of behavioral measurements. New York: John Wiley, 1972.

Evans, M. G. Conceptual and operational problems in the measurement of various aspects of job satisfaction. Journal of Applied Psychology, 1969, 53(2), 93-101.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954. (Second Edition)

Guttman, L. Measurement as structural theory. Psychometrika, 1971, 36(4), 329-347.

Klein, H. A. The world of measurement. New York: Simon and Schuster, 1974.

Krantz, D. H. Measurement structure and psychological laws. Science, 1972, 175(4029), 1427-1435.

Kuhn, T. S. The function of measurement in modern physical science. In H. Woolf (Ed.) Quantification: A history of the meaning of measurement in the natural and social sciences. Indianapolis: Bobbs-Merrill, 1961. Pages 31-63.

Lorge, I., Cronbach, L. J., Scates, D. E., and Tucker, L. The fundamental nature of measurement. In D. N. Jackson and S. Messick (Eds.) Problems in human assessment. New York: McGraw-Hill, 1967. Pages 43-56.

Luce, R. D. What sort of measurement is psychophysical measurement? American Psychologist, 1972, 27(2), 96-106.

McCormick, E. J. Human factors engineering. New York: McGraw-Hill, 1964. Pages 133-138. (Second Edition)

Meister, D. Behavioral foundations of system development. New York: John Wiley, 1976.

Meister, D. Subjective data in human reliability estimates. Proceedings 1978 Annual Reliability and Maintainability Symposium, in press.

Salvendy, G. and Seymour, W. D. Prediction and development of industrial work performance. New York: John Wiley, 1973.

Sechrest, L. Another look at unobtrusive measures: An alternative to what? In R. W. Sinaiko and L. A. Broedling (Eds.) Perspectives on attitude assessment: Surveys and their alternatives. Washington: ONR Technical Report TR-2, August 1975. Pages 103-116.

Stevens, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.) Handbook of experimental psychology. New York: John Wiley, 1951. Pages 1-49.

Webb, E., Campbell, D. T., Schwartz, R. D., and Sechrest, L. Unobtrusive measures: Nonreactive measures in the social sciences. Chicago: Rand McNally, 1966.

Wilks, S. S. Some aspects of quantification in science. In H. Woolf (Ed.) Quantification: A history of the meaning of measurement in the natural and social sciences. Indianapolis: Bobbs-Merrill, 1961. Pages 5-12.

ABOUT THE AUTHOR

Frederick A. Muckler is Program Director, Design of Manned Systems, at the Navy Personnel Research and Development Center. He received his A.B., M.A., and Ph.D. degrees in psychology from the University of Illinois. His initial professional experience was at the Aviation Psychology Laboratory of the University of Illinois. From that organization, he moved to The Martin-Marietta Corporation in Baltimore and subsequently to The Bunker-Ramo Corporation in Los Angeles. In 1966 he was a co-founder and President of Manned Systems Sciences, Inc., in Northridge California. In 1975 he moved to San Diego, and his present position. He has concentrated on human factors in system design. He has taught at UCLA, USC, California State University (Northridge), and the California State University (Los Angeles). From 1964 through 1968, he was Editor of Human Factors, the journal of the Human Factors Society. Currently (1976-1977) he is President of the Human Factors Society.

SIMULATION FOR PERFORMANCE MEASUREMENT

Alice M. Crawford
John F. Brock
Navy Personnel Research and Development Center
San Diego, California 92152

ABSTRACT

Simulator use for performance measurement is discussed. Past research, although limited, is reviewed. Simulation capabilities are discussed in terms of cost effectiveness, presentation of stimuli, and data collection. The issues of fidelity, reliability, and validity as they apply to performance testing with simulators are also discussed. Conclusions about the role of simulation in performance assessment are drawn and recommendations for future R&D are presented.

INTRODUCTION

Ever since Musterberg (1913) tried to measure streetcar operator behavior without benefit of a streetcar, test and measurement researchers have been trying to find out how well a person could perform a job without actually having him do that job. Technology has eliminated the streetcar operator but not the general need for predicting on-the-job performance in off-the-job environments.

Two methodologies have generally been used for performance measurement: (1) pencil and paper measures of job knowledge and personal characteristics, and (2) performance tests of specific job skills. Pickering and Anderson (1976) describe research efforts that indicate very low correlations between pencil and paper measures and actual job performance. These authors point out that performance tests are a superior means of assessment since they can provide diagnostic information and offer "one of the most direct means of determining whether or not individuals are capable of performing critical portions of their jobs" (Pickering & Anderson, p. 3, emphasis in original). While performance tests offer certain advantages, they are considerably more time-consuming and expensive to develop than pencil and paper measures. As a result, assessment by means of performance tests is relatively rare.

The present paper will discuss the use of simulation for performance measurement. Simulation will be used to refer to physical representations of equipment and displays acutally used in the work environment. The orientation will be toward assessment in the military for these general purposes: (1) selection, (2) training, including determination of proficiency of the trainee and detection of deficiencies in the instruction, and (3) personnel management including assessment for promotion, retention, assignment, or evaluation of readiness.

An examination of simulation for performance measurement is motivated by two factors. First, previous research in simulation for training has consistently demonstrated benefits over conventional methodologies (e.g., lower cost). If

simulation techniques could successfully be used for assessment, the advantages of actual performance testing could be realized and some of the inherent problems eliminated. The second factor is that new advances in simulation technology offer even more benefits than have been previously realized. Included here are videodisc, computer-based instructional (CBI) systems, hybrid systems (a combination of real and simulated devices), and knowledge-based computer (KBC) systems. These media have either already been used for simulation in training or are intended for use in the near future. They promise not only lower cost but also increased instructional capabilities over their predecessors. Investigation of all possible uses of these devices is needed; the question of how the devices might be used in performance measurement seems particularly provocative.

Two basic questions seem immediately pertinent to the military: (1) "Can a simulated performance test provide the same measurement as an actual performance test?", and (2) "What are the associated advantages?" Since there has been almost no systematic research in the area, it is not possible to thoroughly answer either question at this time. However, some relevant information is available from preliminary research efforts and from the extensive literature on the use of simulation for training.

The present paper will present the available information and point out areas of required research. It is intended that this presentation will provide the reader with sufficient background to determine whether the topic is worthy of further investigation. As the two questions imply, the authors are concerned only with how well simulation can approximate actual performance testing and the benefits that might accrue. Thus, the more basic issues (such as what should be measured) will be mentioned in passing but are not of present concern. Also, many of the issues associated with simulation for training will surface with the discussion of simulation for performance measurement. These will be discussed with an emphasis on how they might interact with performance measurement.

SIMULATION CAPABILITIES

Cost Effectiveness

In most cases, the major advantage of simulation is lower cost than on-the-job operations. Simulation saves wear and tear on expensive equipment and saves fuel where vehicle simulation is concerned. This is particularly important for training in which repeated trials are usually necessary.

Videodisc, KBC systems, and CBI systems are especially inexpensive due to their general purpose nature. These systems are capable of presenting simulations of numerous operational situations or equipments at a single computer terminal; the operator need only access the appropriate software. Additionally, they are designed for extensive interaction with the operator. They present stimuli (e.g., a graphic display of a piece of equipment with text to guide operations), record and evaluate input (the operator's simulated performance is sensed by the computer through a positional input device such as light pen or touch panel), and provide feedback. Thus, the need for an instructor or observer is obviated. (See Crawford, Hurlock, Padilla, & Sassano, 1976, for a cost analysis of a CBI system simulation for experimental training).

It is estimated that videodisc, which will be generally available in about a year, will be the least expensive of the new technologies. With this medium it is possible to present audio along with color, photographed visuals, which can be

still or dynamic. Up to 54,000 frames can be stored on one disc, which is the size of a standard long playing record, and the entire unit is correspondingly small as compared to other systems. The main disadvantage of videodisc is that the interactive capabilities are somewhat fewer than those of the CBI or KBC systems.

KBC systems are similar to CBI in terms of the kind of presentations available (e.g., still or dynamic computer graphics). They differ in that KBC seeks to represent knowledge structures. This is an advance in technology; however, KBC systems are still in the early development stages of software design, which puts them considerably behind CBI systems. Knowledge structures, once developed, can handle a wide variety of individual differences without the need for extensive preprogramming of every possible input. Eventually this will mean extreme ease and low cost of development of simulation materials.

While these devices have been exclusively used for training, there are no apparent limitations on the context within which they can be used. Utilization for performance measurement would only require programming the simulation displays (as for training) and the insertion of branching sequences and text deemed appropriate for the testing situation. Performance has been measured in this manner (e.g., Trollip, 1977), but not for the purpose of investigating the relationship between simulated and actual performance testing.

The programming languages used for CBI systems have evolved to the point where they are relatively easy to use. The tutor language on the PLATO IV instructional system, for example, is referred to as an "authoring" language, implying that the person who designed the materials can probably do the programming without extensive experience with computers. (See Hurlock & Slough, 1976, for estimates of required developmental time.) Research should bring about similar developments for the other systems, which will mean ease of use and, therefore, lower cost, for a wide variety of media. In general, the cost of computers is decreasing as technological sophistication increases.

While the computer-based, two-dimensional simulation systems are interesting and may offer unique advantages for performance measurement--as will be seen in the following sections--"simulation" in the present context also refers to three-dimensional and hybrid simulators. Three-dimensional systems would become too expensive if an attempt was made to implement the kind of computer support required for automatic stimulus presentation and data collection as already exists in the computer-based systems. However, the two-dimensional systems will probably not be able to replicate all tasks with sufficient accuracy because of their lack of physical fidelity. The hybrid systems offer some of the advantages of each type, but it is not clear at the present time whether a little of each system will be better than the total capabilities of either. The cost and capability tradeoffs for all simulation systems will have to be determined by future work.

Presentation of Stimuli

An important characteristic of simulation for performance measurement is that representations of events which are normally unavailable under actual conditions may be presented. This permits controlled measures of performance in emergency situations, repair of malfunctioning equipment, and other situations which are

otherwise dangerous or not normally accessible. Additionally, frequency of these events can be presented without real-time constraints, which should facilitate more accurate measurement. Engagement simulation (Gorman, 1976), in which personnel participate in simulated battle conditions, is a good example of how simulation may be used in place of a dangerous real-world situation.

Another feature of simulation is that varying levels of difficulty of a task, or isolated parts of a task, may be presented. Literature reviews show that, when performance goals are too difficult, a person may give up or produce invalid data to make it appear that the goal has been reached (Porter, Lawler, & Hackman, 1975). In this context, it seems possible that by varying difficulty levels or by isolating a task segment, the quality of the information available for selection decisions could be enhanced.

Data Collection

The advantages of automated data collection have been emphasized by many investigators (e.g., Vreuls & Obermayer, 1971). The new computer-based technologies should prove capable of performing this function for large amounts of data at relatively low cost. Additionally, where performance measures can be objectively defined, the computer can concomitantly evaluate the data. With ambiguous measures, performance data can be stored for postperformance evaluation by several observers.

In addition to how much can be measured with simulation, it is worthwhile to note what can be measured. The topic of process vs. product measurement has been thoroughly discussed and the pros and cons of each side weighed (e.g., Shriver & Foley, 1974). Many researchers, such as Pamitz and Olivo (1971), have concluded that both types of measurement are often necessary for a thorough evaluation of performance; the present authors agree.

For selection and training purposes, detailed information is needed to reveal skill levels or areas of deficiency, and these can only be determined through process measurement. In the case of personnel management, it seems that only the product is of interest. This position makes intuitive sense in that, with personnel who are presumably skilled, it is desirable to know that they can do the job, not how they do the job. However, research has shown that measurement of one type may lead to deemphasis on the other, which, in turn, may produce dysfunctional consequences for the organization (Porter, et al., 1975). For example, emphasis on product measurement may cause short-term results maximization with failure to perform certain important functions. On the other hand, measurement of process alone may encourage rigid bureaucratic behavior with a possible lack of innovative accomplishments.

While job characteristics may greatly influence what aspects of particular tasks (i.e., process or product) will get measured, simulated performance measurement can accomplish both since it is relatively free of the logistical constraints that have inhibited this in the past (e.g., the impossibility of human monitoring of every aspect of a complex task or sufficient computer memory to record it all).

The subject of improved evaluation capabilities is also pertinent to mention at this time. This refers to unique, and presumably more accurate, ways to measure

process or product that weren't previously available. As researchers continue to identify and quantify optimum solutions to problems, there will be a technology to support the implementation of these new measurement techniques to assess problem solving strategies. Trollip (1977), for example, has utilized the considerable capacity of PLATO IV to implement continuous performance measurement in the simulated flight of holding patterns.

Hyatt and Deberg (1976) discuss an energy maneuverability (EM) index for measurement of air combat maneuvering (ACM). The EM data serve as the nucleus of an algorithm with three major aims: (1) predicting success in air-to-air combat, (2) measuring process in ACM training, and (3) improving cockpit displays of ACM. The authors point out that this technique could be carried out on real equipment given sufficient computer support. This illustrates another advantage of simulation. It is rarely feasible to add computer support to an actual device (such as an aircraft) for performance testing. Specifically, costs, storage requirements, and maintenance problems would accelerate tremendously.

THE ISSUES

Having seen the advantageous features of simulation and how they apply to performance measurement, it is appropriate to turn next to the issues involved in determining the relationship between simulated and actual performance measurement. Fidelity, considered to be of major importance, will be discussed first. Following this, reliability and validity will be considered. In the present context, these later concepts become somewhat ambiguous and will be mentioned primarily as background for future research. Upon clarification of these issues, potential directions for future research will be presented.

Fidelity

Appearance and functional fidelity have been major concerns in simulation research for years. Simplistically stated, researchers want to know how much a simulator must look and act like the real thing to ensure positive transfer of training.

For a long time it was the accepted belief that there was a direct positive relationship between high fidelity (particularly appearance fidelity) and high transfer of training. Eventually, researchers began to suggest that the quality of the instructional context might have more bearing on transfer than physical similarity between simulators and real operational equipment and processes (e.g., Micheli, 1972). In fact, recent research has indicated that this may often be the case (e.g., Caro, 1973).

This trend toward stressing instructional content is still present today. There is much concern with the manipulation of learning variables such as learner control of course content (Lahey and Coady, 1977), and the Zeitgeist seems to be swinging more toward hardware design being driven by training concerns.

The problem of fidelity for simulated performance measurement seems, at least superficially, to be quite different from the training concerns described above. However, as was suggested in the section on the capabilities of simulation, the

problems associated with simulated performance measurement may not be very different from the simulation problems dealt with by the training community. At the very least, previous training research should provide guidance for this new area.

Starting from scratch, it will be necessary to ask the following: "What are the properties of fidelity which must be considered in the design of simulation for performance measurement to ensure that it is an accurate representation of an actual performance measure?" Since this question requires nothing less than an algorithm that would specify the level of fidelity of simulation to be used to assess each given performance for each given purpose, it will not be answered until a considerable amount of research has been undertaken. It should be a helpful start, however, to discuss some of the relevant research findings.

In troubleshooting tab tests, Crowder, Morrison, and Demaree (1954) found correlations between their tests and on-the-job performance to range between .12 and .16. In a similar study, Steinemann (1966) found correlations ranging from -.50 to .14. He commented:

> In the actual task, students were reluctant to unsolder or disconnect components from the chassis, but in the simulated task, where parts replacement required virtually no effort, students too often resorted to parts replacement in an effort to solve the problem. (pp. 10-11).

In this case, it appeared that the simulated situation was not realistic enough to be an accurate representation of on-job performance. Additional fidelity was needed to account for the difficulty required to do the actual task.

In another study (Baron & Williges, 1975), simulation for training driving skills produced performance measurement data that were found to be of questionable validity by the researchers. These authors attribute this, in part, to a lack of fidelity caused by an open-loop system. Specifically, simulator drivers did not receive realistic proprioceptive, visual, and audio feedback as a result of their performance.

Abrams, Schow, and Riedel (1974) report on the design of a welding simulator for training purposes that yielded correlations of .68 to .73 between simulated and actual welding performance. The simulator required a person to track a target that moved in horizontal and lateral dimensions like a welder would actually move his welding rod. The person tracked with a rod that shortened much as it would in actual welding. In the training mode, the student received audio and visual cues when he was off the track, when the angle of the rod exceeded limits, or when the rod was too close to or too far away from the track. All errors were recorded. For testing, the augmented cues could be dropped but the scoring kept.

The success of the welding simulator may have been due to its high functional fidelity, or it may have been a result of the fact that welding is a task that can be clearly defined and, therefore, preprogrammed without ambiguity.

On a more complex level, Koonce (1974) had pilots perform a mission in a simulator under one of three conditions of motion (sustained-linear, washout, or no motion)

and then in an actual aircraft under similar conditions. Koonce concluded that his results had shown that the proficiency of aircraft pilots can be predicted to a high degree from ground-based performance using simulators. He also noted that greater prediction could be obtained using sustained motion as compared to the other conditions. The author also concluded that there were very high observer-observer reliabilities (R=.77 to .97) and attributed this to well-defined, easy to follow measurement scales.

Based on the studies reviewed, it appears that a certain level of functional fidelity is as important in simulated performance measurement as it is in training. There were high correlations between simulated and actual performance measures where there were clear visual and/or proprioceptive cues and feedback. Low correlations were found where these cues were absent or where the difficulty level was unrealistic. Additionally, the component of well-defined measures cannot be overlooked since this factor was also present in the studies which had high functional fidelity. Whether these findings represent future research trends remains to be demonstrated.

Reliability and Validity

Discussions of the reliability and validity of simulated performance measures are only appropriate after the critical elements of actual performance are determined. The central theoretical issue concerns what is a valid and representative work sample. That is, what are the essential performance measures to assess on the job performance.

Assuming that a high correspondence between actual performance measures and effective job performance has been established, one can attempt to assess the reliability and validity of simulated performance measures. While some global statements can be made, the following discussion should indicate that the meaning of these concepts is altered by the unique nature of simulated measurement. It probably will be necessary, eventually, to reevaluate these issues in the context of new research.

Test-retest reliability of performance measurement should improve when simulation is used in place of real operational conditions. Pickering and Anderson (1976) have noted that, when job experts or instructors in the military are doing the testing, they often don't maintain standardized testing procedures. They are likely to coach and to give feedback as if they were operating in a training mode. Given the capability of the computer to perform standardized testing functions, error variance generated from instructor behaviors, such as inconsistent presentation of materials and feedback, should be eliminated.

Another factor influencing reliability is the complexity of the process being measured. Koonce (1974), for example, found that inter-observer correlations used in assessment of flight proficiency were influenced by several factors: (1) instrument-referenced items were apparently easier to judge than points outside the vehicle, (2) some instruments were easier to read than others, and (3) recording errors were sometimes caused by a lack of training in the observers. All of these problems could be eliminated through simulation in which a computer is programmed to precisely record every step in the performance process.

Knoop and Welde (1973) found that reliability varied as a function of aircraft and environmental conditions. Again, part of this variance would be eliminated when a computer performed the observer function. However, simulation could also provide an added benefit here -- different environmental conditions could be programmed into a simulator, and the operators could then be tested at each level.

The unique nature of the psychometric problems associated with simulated performance measurement become more clear when internal consistency measures of reliability are considered. With operational skills there is often only one product and one process measure for a given task. Therefore, there is no way to obtain similar measures and determine intermeasure reliability.

If different aspects of a process are measured, the issue appears to concern validity more than reliability. Specifically, the question then becomes, "Do different aspects of the task represent the same content area or the same underlying construct?" In fact, it is the position of the present authors that, for simulated performance testing, construct validity is the crucial issue and reliability is probably not an important issue.

Types of validity for simulated performance measures tend to overlap, but there are three very broad categories that can be conceptualized with each type corresponding to the stated purpose of the assessment. The first is predictive validity for selection. For example, it would be of interest to determine if a simulated performance measure from a person completing a Navy "A" school accurately predicts later on-the-job performance. If valid, such measures could be used to differentially select individuals for more critical organizational sites. The second area is content validity for training; that is, assurance is needed that representative content areas of a job have been sufficiently tapped to enable effective training of a person. The third categorization is construct validity for personnel management. Here, the simulated measurement does not have to be representative of all tasks performed in a job, but must be an accurate measure of the underlying construct that represents effective job performance.

As stated earlier, construct validity may be the most important issue. If there is construct validity, the categorizations become indistinguishable since construct validity also assures that there is predictive validity. Simply stated, if a measure shows effective job performance, it obviously predicts that the job can be done. Thus, the most important issue is the determination of what provides the best measures of effectiveness on the job.

Adequate task analysis may provide a good start on some of these problems. Klein (1976) suggests that the maintenance and operation of complex systems may not be amenable to traditional task analytic techniques. However, there is preliminary evidence that a logical, hierarchical analysis of required tasks will, in fact, produce the behavioral data necessary to begin specification of the design characteristics for a simulator (Brock, 1976; Malone, DeLong, Farris, & Krumm, 1976).

Additional research is needed to specifically determine an analytic process that will isolate those qualities of job performance that are most crucial for the purpose described here. Rundquist (1977) has proposed a program for comparative analysis of all basic task analysis methods; however, there are no data to support the efficacy of his approach.

CONCLUSIONS AND RECOMMENDATIONS

Cream (1976) has noted that, in assessing performance, "we have often chosen to measure that which we could measure, rather than that which should have been measured" (p.26, emphasis in original). For tasks involving man-machine interactions, simulation is a potential solution to the problem. Given a simulation medium with sufficient computer support, capabilities for measures of interest can be specified in the hardware and software design, and the entire sequence of a person's performance can be recorded and evaluated. Also, the measurement process achieves higher levels of standardization and objectivity.

All types of simulation can offer the capability for measurement of situations that would be dangerous or unavailable during on-the-job performance, and most of them cost less than assessment under actual conditions. The new computer-based technologies, such as videodisc, may offer additional benefits such as (1) capability for measurement of tasks at varying levels of difficulty or environmental conditions, (2) availability of product and process measures on tasks that would be too involved for a human observer to record, and (3) capacity for unique and improved evaluation of complex performances.

Characteristics of the new simulation media could be conducive to military performance testing in assessment centers. An assessment center at each military installation could provide a convenient method by which to locally assess performance for selection, training, or management purposes. Given the small physical space requirements and general purpose features of these devices, a wide range of testing could be carried out at each center for a large number of persons.

In spite of the considerable number of advantages that simulation could offer for performance measurement, it is not clear at this time whether it offers a preferable alternative to actual performance measurement. Considerable research will be necessary to make this determination. As implied earlier, research is needed to assess the degree to which simulated performance measures are representative of actual performance measures. Likewise, the cost effectiveness of such simulation must also be determined.

In general, it can be said that research is needed to determine the correlations between any type of simulation and the on-job performance it corresponds to. This information is needed for a representative hierarchy of task behaviors. These data could be collected by testing performance of different tasks at systematically degraded levels of fidelity, thus (1) providing information that could be used to derive an algorithm, and (2) specifying when and where to use different types of simulation. This is particularly important since the simulation systems that appear to have the most benefits (i.e., the two-dimensional systems) also have the lowest fidelity. Such an algorithm is also needed by researchers involved in the use of simulation for training, and work in the area is in progress. (e.g., Miller, McAleese, Erickson, Klein, and Boff)[1] These continued developments should also prove helpful to investigators interested in simulation for performance measurement.

---

[1] Miller, L. A., McAleese, K. J., Erickson, J. M., Klein, G. A., and Boff, V. R. Draft: Training device design guide, the use of training requirements in simulation design. Prepared for Air Force Human Resources Laboratory, June 1977. Limited Distribution.

More detailed research is needed to evaluate the specific characteristics of different types of simulation as well as various issues related to any type of simulation. Recommended areas of investigation are the following:

1. Videodisc is among the fanciest of the low cost media; however, operational capabilities such as motion and color may not outweight its lack of a capability for extensive man-machine interaction. The tradeoffs should be evaluated.

2. CBI systems have been shown to be effective for training certain types of tasks; they would probably also be effective for measuring certain kinds of performance. As will be necessary for all computer-based simulation systems, there is a need to determine where two-dimensional simulation is no longer sufficient. An efficient approach would probably be to look at this factor in light of characteristics of the subject population under consideration. For example, the performance of experienced pilots on a part-task, two-dimensional simulation measure would probably not be comparable to performance in a complex aircraft.

3. KBC systems have fascinating implications for training, which may be beneficial for assessment. They have capability for adaptation to individual skill levels, which may prove useful for a detailed analysis of deficiencies in performance. This seems worthy of investigation once the state-of-the art reaches an appropriate level.

4. The hybrid simulator, offering both a sufficient amount of fidelity and many of the capabilities of the computer-based systems, may potentially provide the best source of simulated performance measurement. Nonetheless, considerable empirical work remains to be done in this area for these simulators to reach their potential.

5. Situational simulation should be evaluated over a wide range of different tasks. In this case, a scenario is presented to personnel which requires verbal or written report of decisions and actions which should occur under actual conditions. This methodology could prove effective for either simple tasks or abstract tasks.

6. Reliability and validity issues and techniques should be reevaluated within the context of the unique characteristics of simulation and performance measurement. This should include evaluation and refinement of task analysis methodologies.

The present paper has discussed the potential benefits of using simulation for performance measurement and pointed out areas in which research is needed. It seems appropriate to add this final reminder, that simulation is only a tool that must be ultimately evaluated within the context of how all personnel involved use it and are affected by it. Thus, all simulation research must account for the human component from the perspective of both management and the persons being assessed. Failure to do so will result in a sophisticated technology which cannot be successfully implemented within the dynamics of military organizations.

REFERENCES

Abrams, M. L., Schow, H. B., & Riedel, J. A. Acquisition of a psychomotor skill using simulated-task, augmented feedback (Evaluation of a welding training simulator). (NPRDC Tech. Rep. 75-13). San Diego: Navy Personnel Research and Development Center, October 1974.

Baron, M. L., & Williges, R. C. Transfer effectiveness of a driving simulator. Human Factors, 1975, 17(1), 71-80.

Brock, J. F. Development of a task category system for the design of air crew training. Paper presented at The Fifth Psychology in the Air Force Symposium, United States Air Force Academy, Colorado, April 8-10, 1976.

Caro, P. W. Aircraft simulators and pilot training. Human Factors, 1973, 15,(6), 502-509.

Crawford, A. M., Hurlock, R. E., Padilla, R., and Sassano, A. Low cost part task training using interactive computer graphics for simulation of operational equipment (NPRDC Tech. Rep. 76TQ-46). San Diego: Navy Personnel Research and Development Center, September 1976. (AD-A029 540/2WS)

Cream, B. W. Training requirements and simulator utilization. In First International Learning Technology Congress and Exposition on Applied Learning Technology Proceedings Volume IV: Future of Simulators in Skills Training. Washington, D.C., July 21-23, 1976, 24-27.

Crowder, N., Morrison, E. J., and Demaree, R. G. Proficiency of Q-24 radar mechanics: VI. Analysis of intercorrelations of measures (AFPTRC-TR-54-127). Lackland AFB, TX: Air Force Personnel and Training Research Center, 1954.

Gorman, P. F. Engagement simulation. In First International Learning Technology Congress and Exposition on Applied Learning Technology Proceedings Volume IV: Future of Simulators in Skills Training. Washington, D.C., July 21-23, 1976, 136-141.

Hurlock, R. E., and Slough, D. A. Experimental evaluation of Plato IV technology: Final report. (NPRDC Tech. Rep. 76TQ-44). San Diego: Navy Personnel Research and Development Center, August 1976.

Hyatt, C. J., and DeBerg, O. H. Advanced performance measuring systems for flight simulators. In First International Learning Technology Congress and Exposition on Applied Learning Technology Proceedings, Volume IV: Future of Simulators in Skills Training. Washington, D. C., July 21-23, 1976, 68-71.

Klein, G. A. Problems and opportunities in deriving training requirements for the design and utilization of simulators. In First International Learning Technology Congress and Exposition on Applied Learning Technology Proceedings Volume IV: Future of Simulators in Skills Training. Washington, D.C., July 21-23, 1976, 142-146.

Knoop, P. A., and Welde, W. L. Automated pilot performance assessment in the T-37: A feasibility study. (AFHRL TR 72-6). Wright-Patterson AFB, Ohio: Advanced Systems Division, Air Force Human Resources Laboratory, April 1973.

Koonce, J. M. Effects of ground-based aircraft simulator motion conditions upon prediction of pilot proficiency. (ARL-74-5/AFOSR-74-3). Savoy: University of Illinois at Urbana, Champaign, Institute of Aviation, Aviation Research Laboratory, April 1974.

Lahey, G. L., and Coady, J. D. Student response to guided computer-based instruction (NPRDC technical report). San Diego: Navy Personnel Research and Development Center, in preparation.

Malone, T. B., DeLong, J. L., Farris, R., & Krumm, R. L. Advanced concepts of Naval engineering maintenance training (Vol. I of II). (NAVTRA EQUIP CEN N-61339-74-C-0151). Orlando: Naval Training Equipment Center, May 1976.

Micheli, G. Analysis of the transfer of training, substitution, and fidelity of simulation of training equipment (TAEG Report 2). Orlando: Naval Training Equipment Center, Training Analysis and Evaluation Group, 1972.

Munsterberg, H. Psychology and industrial efficiency. New York: Houghton Mifflin, 1913.

Pauitz, A., and Olivo, C. T. National occupational competency testing project. A consortium for occupational competency testing of trade and industrial technical teachers. Phase I: Planning - organizing - pilot testing. Handbook for developing and administering occupational competency tests (Vol. 3). Washington, D.C.: Office of Education (DHEW), February 1971.

Pickering, E. J., and Anderson, A. V. Measurement of job-performance capabilities. (NPRDC Tech. Rep. 77-6) San Diego: Navy Personnel Research and Development Center, December 1976.

Portler, L. W., Lawler, E. E., III, & Hackman, J. R. Behavior in Organizations. New York: McGraw-Hill, 1975.

Rundquist, E. A. Developing criteria for judging the results of a job analysis for training design purposes (NPRDC technical report). San Diego: Navy Personnel Research and Development Center, in preparation.

Shriver, E. L., and Foley, J. P.,Jr. Evaluating maintenance performance: The development and tryout of criterion referenced job task performance tests for electronic maintenance. (AFHRL-TR-74-57(II), Part I) Wright-Patterson Air Force Base, Ohio: Advanced Systems Division, Air Force Human Resources Laboratory, September 1974.

Steinemann, J. H. Comparison of performance on analogous simulated and actual trouble shooting tasks. (PRA SRM 67-1). San Diego: Naval Personnel Research Activity, July 1966.

Trollip, S. R. An evaluation of a computer-based flight procedures trainer (ARL-77-1/AFOSR-77-1). Savoy: University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Research Laboratory, February 1977.

Vreuls, D., and Obermayer, R. W. Emerging developments in flight training performance measurements. Naval Training Device Center's 25th Anniversary Commemorative Technical Journal, November 1971, 199-210.

ABOUT THE AUTHORS

Alice M. Crawford has worked at the Navy Personnel Research and Development Center as a Research Psychologist for four years. She is assigned to the Training Technologies Department where she has been involved in the experimental evaluation of the PLATO IV instructional system, computer-based simulation for S-3A operations and tactical training, and computer-managed instruction. She is currently coordinating the development of an encyclopedia of articles intended for use in curriculum design. Ms. Crawford received her M.A. degree in Experimental Psychology from San Diego State University in 1973.

John F. Brock has been an Education Specialist at the Navy Personnel Research and Development Center and its predecessors for 10 years. Prior to that, Mr. Brock served as a Navy officer on board USS PERKINS (DD-887), USS WRIGHT (CC-2) and the Fleet Anti-Air Warfare Training Center, San Diego. He has conducted research into instructional systems design, programmed instruction, shipboard training systems, engineering maintenance training, aircrew training development and most recently, has begun a R&D program on developing a Life Cycle Costing Model for instructional systems. Mr. Brock has done graduate work in experimental psychology at San Diego State University. He is a member of the Human Factors Society, The American Education Research Association, The National Society for Performance and Instruction, and the Society for Applied Learning Technology.

MEASUREMENT OF PRODUCTIVITY ENHANCEMENT: EVALUATING A PERFORMANCE-CONTINGENT REWARD SYSTEM THAT USES ECONOMIC INCENTIVES

Gene E. Bretton
Steven L. Dockstader
Delbert M. Nebeker
E. Chandler Shumate
Navy Personnel Research and Development Center
San Diego, California 92152

ABSTRACT

The cost/effectiveness, cost-savings projections, and related issues of a Performance-Contingent Reward System (PCRS) that uses economic incentives to enhance productivity were evaluated. The PCRS was tested on civil-service personnel functioning as data transcribers in a Management Information System Department of a large Navy shipyard on the West Coast. Evaluation of the PCRS was primarily from the following perspectives: (1) cost/effectiveness of the proposed PCRS relative to former production conditions at the test site, (2) issues involving generalizability of the test-site results to other federal sites with substantial concentrations of data transcribers, and (3) projections of PCRS-induced cost savings in terms of specified (a) outyears, (b) levels of aggregation of data transcribers, and (c) levels of generalizability of test-site results.

It was recommended that managers having control over sites with large numbers of civil-service data transcribers evaluate the results of the PCRS field test from the perspective of possible implementation. Such managers should give special attention, of course, to issues underlying the generalizability of test-site results to sites under their control.

PRECEDING PAGE BLANK-NOT FILMED

INTRODUCTION

Problem

Unless performance enhancement can be accurately measured and evaluated, its true value cannot be assessed. Determining the appropriate procedures that will permit accurate measurement and evaluation is, therefore, an integral part of well-planned productivity enhancement efforts.

Purpose

The primary purpose of this paper is to explore some interrelated R&D and operational issues of implementing a Performance-Contingent Reward System (PCRS) that uses economic incentives. The PCRS was designed, in part, to enhance productivity and morale of personnel in Federal employment settings. To determine if the PCRS is effective in achieving these objectives requires, as a minimum, the measurement and evaluation of field-test results in terms of both the validity and economic value of the enhancement.

1. Validity. The issue of scientific validity has multiple dimensions, which may be broadly defined as follow:[1]

   a. Statistical Conclusion Validity, which deals with whether or not a presumed cause and effect covary beyond the expecatations of chance.

   b. Internal Validity, which addresses the degree of causality that can be correctly inferred from observed relationships.

   c. Construct Validity, which deals with the fidelity with which theoretical relationships are operationalized.

   d. External Validity, which refers to the generalizability of causal relationships across different persons, settings, or times. The ways in which these dimensions relate to various aspects of the PCRS field test will be examined in subsequent sections of this paper.

2. Economic Value. The value of productivity enhancement can be measured and evaluated in such diverse units as physical outputs, "utility", or dollars. Although each of these units has advantages and disadvantages, most managers and other policy-makers appreciate the increased interpretability of productivity enhancement efforts that can be evaluated in economic terms. The overall benefits of converting results of productivity enhancement into dollars, if feasible, are obvious and need no elaboration.

There are many interrelationships between the validity and the economic value of productivity enhancement. Some of these interrelationships are explored in the economic evaluation of the PCRS, which was conducted primarily from the following perspectives:

---

[1] For those wishing a comprehensive and in-depth analysis of validity issues--especially from the perspective of planning, conducting, and evaluating field research--the chapter by Cook and Campbell (1976) in The Handbook of Industrial and Organizational Psychology is excellent.

1. Cost/effectiveness of the PCRS relative to former production conditions at the test site.

2. Issues involving generalizability of the test-site results to other federal sites.

3. Projections of PCRS-induced cost savings in terms of specified (a) outyears, (b) levels of aggregation of designated type of personnel, and (c) levels of generalizability.

In evaluating the generalizability of the test-site results to other federal sites (2 above), an appropriate balance of emphasis on financial, administrative, behavioral, and other issues dealing with PCRS implementation was required. Thus, to ensure that financial issues received adequate--but not excessive--emphasis in addressing the diverse implementation issues, data on financial issues were presented in a form designed to meet two objectives. First, it had to be meaningful to all managers and staff personnel in the federal community, some of whom are not involved with financial analysis. Second, it should enable the technical staff of each implementation site to extend analysis of the data, if necessary, toward meeting the specialized needs of their respective managers (who will be functioning at various hierarchical levels and in diverse organizational settings). Thus, the data are presented in a manner amenable, for example, to (1) adjusting cost-savings projections of constant dollars into dollars reflecting anticipated inflation, (2) converting future values of cost savings into their net present values, (3) doing sensitivity analysis on important parameters affecting the cost savings, and (4) calculating appropriate savings/investment ratios and investment payback periods.

Background

State of the Art

The general status of work incentives and related issues is well stated by Belcher (1974):

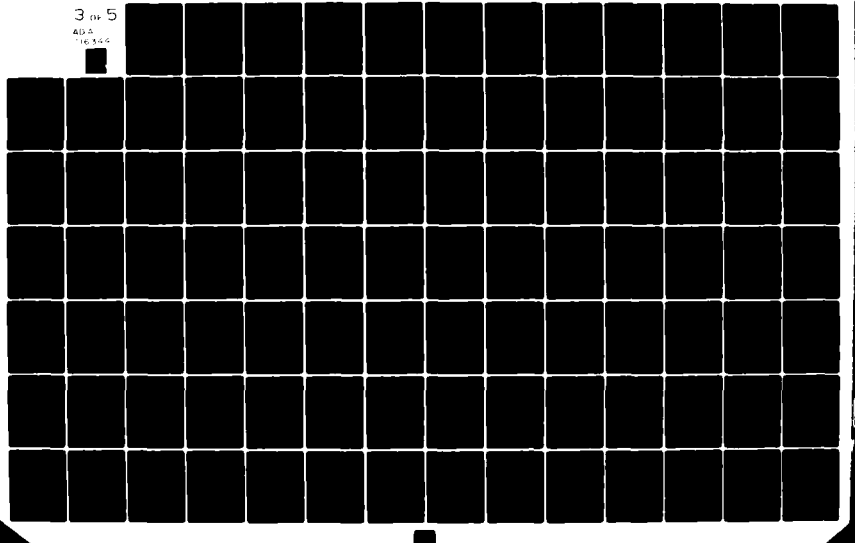
> Incentive plans are controversial. Opponents range from those who oppose the idea of performance rewards on the grounds that performance is a function of the organization of work and management practices rather than employee effort, to those who oppose incentive plans on the grounds that they don't work and cause more problems than they solve. The decline, perhaps the disappearance, of incentive plans is often predicted.
>
> Proponents of incentive plans often believe that a 'fair day's work' is not normally attainable in the absence of an incentive plan because the workers produce only about 50 to 60 percent of the output attained by incentive workers. Although they admit that some incentive plans malfunction, they insist that this is usually due to poor installation and maintenance rather than the concept of incentive. (p. 300)

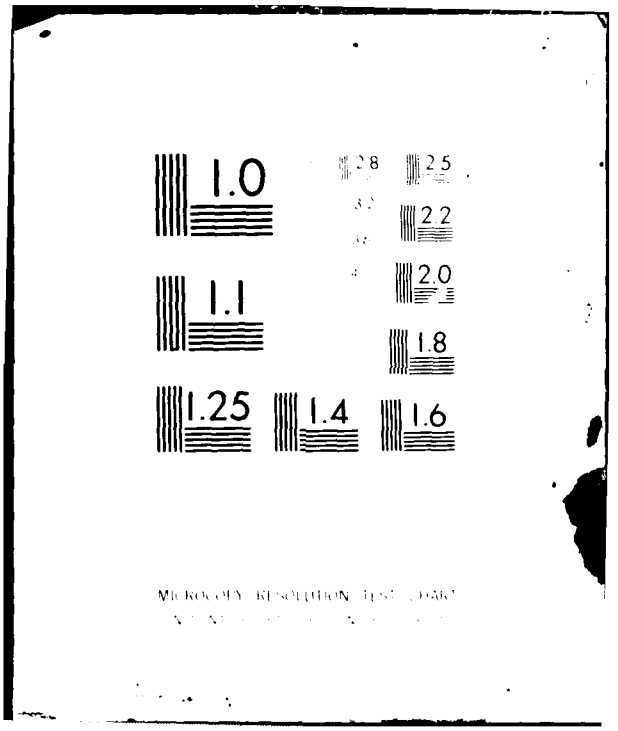
The quotation amply illustrates the deep-seated nature of the controversy surrounding the effectiveness of work incentives. Most managers, wage and salary administrators, labor economists, industrial engineers, and behavioral scientists are well aware that the effectiveness of incentive plans is con-

AD-A116 344 NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 5/9
SYMPOSIUM PROCEEDINGS: PRODUCTIVITY ENHANCEMENT: PERSONNEL PERF--ETC(U)
1977 L T POPE, D MEISTER

UNCLASSIFIED NL

3 OF 5
AD A
116 344




1.0
2.8
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

tingent on the overall situation in which they are applied. Several important issues that must be addressed in any meaningful evaluation of incentive effectiveness are the following:

1. Are the incentives distributed on an individual, group, or plant-wide basis?

2. Is the work machine-paced, or essentially under worker control?

3. Are the employees primarily managerial/professional personnel or clerks and machine operators?

4. Are the incentives awarded for increased effort, performance, or what?

5. What is the basic climate in the work site in terms of, for example, degree of decision-making participation, union/management relations, and the collective trust that nonmanagement personnel have in their management?

Given that effectiveness of work incentives is contingent on the interrelationships of several complex issues, discussion of how the PCRS relates to these issues will be deferred until the basic procedure of the PCRS has been described and its test results evaluated. For those, however, who wish a fairly comprehensive review of this area from the following substantive perspectives, these references are helpful: Wage and Salary Administration (Belcher, 1974); Industrial Engineering (Fein, 1971); Industrial/Organizational Psychology (Lawler, 1971); and Labor Economics (Perlman, 1969).

APPROACH

Hypotheses

The field study of the PCRS tested the following hypotheses:

1. Implementation of the PCRS will substantially reduce the production cost of data-transcribing activities relative to costs associated with former production conditions at the test site.

2. Implementation of the PCRS will not diminish the production effectiveness of data-transcribing activities while the cost are being reduced.

These hypotheses are directly implied by the "Fixed Effectiveness" mode of cost/effectiveness analysis (Fabrycky & Thuesen, 1974). Simply put, this means that when the effectiveness of each competing alternative is equivalent to all others, the preferred alternative can be chosen on the basis of cost.

Variables

Independent Variables

The independent variable is the presence or absence of the PCRS. Individual "elements" of the PCRS will not be separately related to the various dependent variables. Only the PCRS as a total system will be tested.

Dependent Variables

The dependent variable for measuring "cost" in this study of cost/effectiveness is cost-per-keystroke. In contrast, the dependent variables for measuring "effectiveness" are (1) level of production, (2) efficiency of production, and (3) quality of the production process--as defined in terms of workload backlog and, separately, overtime man-hours used.

Sample

The sample is composed of 17 female civil-service data transcribers selected from three shifts in the Management Information System Department of a large Navy shipyard on the West Coast. Of the 26 data transcribers available, the 17 were chosen because (1) they were fully qualified, as opposed to being in training status, and (2) comparative information was available on each data transcriber for a specified time interval (13 weeks) before the field test began, against which the trial-period results could be contrasted.

It is important to note that the 17 data transcribers constituting the sample in both the baseline period and the trial period are not merely equivalent in size, but identical in terms of personnel involved.

Measures

Recharge Rate (RR)

This is the most important cost parameter used in this study. The

parameter is a multiple-component cost figure that represents the overall cost of the data-transcribing operation. (Derivation of the RR is described in Appendix A.) The level of the RR is adjusted periodically by the controllership department of the test site to ensure the hourly cost per data transcriber is kept as current as possible. Components of the RR, and their present levels, are as follows:

1.	Salary	$4.04
*2.	Acceleration	$1.31
3.	Supervision	$1.59
4.	Machines	$0.38
5.	Overhead (General & Administrative)	$3.50
	Total	$10.82

* Represents government share of leave, pension, and other benefits.

Keystrokes

The number of keystrokes represents the combined print-and-verify activities of the data transcribers. Keystrokes are tabulated for each data transcriber directly by the machine on which the data transcribing is done.

Regular Man-Hours

There are three kinds of man-hours to be considered:

1. Paycard manhours--the basis on which data transcribers receive their regular salary.

2. Assigned-machine-time man-hours--the number of man-hours a particular data transcriber is assigned to operate a specified machine.

3. Actual-machine-time man-hours--the number of man-hours actually spent operating the machine to which assigned. The interrelationships of these different kinds of man-hours are described in Appendix B, which details the PCRS-bonus computation.

Overtime Man-Hours

The number of man-hours permitted by site management to be paid at the overtime rate to keep workload backlog within acceptable limits.

Backlog

This is measured by the average number of batches of work remaining undone, as calculated on a daily basis and then averaged over weekly periods.

A batch is a set of tasks that is fairly homogeneous in content but has some variation in size. Thus, "batch" is a meaningful unit only when randomized across considerable periods of time--such as the total number of work days for the base period versus the trial period in this study: 63 and 64 days, respectively.

Data Sources

Data on the measures just described were collected from sources presumed to vary considerably in ability to reflect the PCRS impact on various site activities during the trial period. Toward ensuring that the data collected were the best available, the sources were used in the following priority:

1. Site Documents. Most of the extremely critical data regarding keystroke production, man-hours, etc., were taken from computer printouts that had summarized data directly from source documents. This was almost the sole source for the important data used in calculating the production cost-savings.

2. Estimates from Site Management and Staff. When site documents were not available for required data, estimates were solicited from appropriate managers and qualified staff personnel.

3. Records of Research Team. Sometimes neither site documents nor site personnel were the best sources for required data. In these instances, appropriate data were recorded or derived by the research team. An example would be the determination of total man-hours required by the research team from various types of site employees for various purposes.

4. Composite Source. Regarding some issues, none of the sources above was capable of providing adequate information by itself. One such issue dealt with determining possible "hidden" set-up costs which might have been incurred, but which were unrecorded and not otherwise readily accountable. Thus, representatives from the site's accounting department and the research team jointly decided to prorate such possible costs from the basis of known set-up costs.

Procedure

Brief summaries of essential procedural steps of the PCRS are described here to illustrate that it is not merely an administrative program for dispensing incentives--but a major organizational modification that has deep impacts on several important dimensions of the work process, and which requires major changes in work roles of workers and supervisors alike.

Procedural requirements of the PCRS will be discussed in five phases: (1) Preliminary Issues, Objectives, and Activities, (2) Development, Administration, and Testing, (3) Preliminary Evaluation, (4) Modification and Maintenance, and (5) Full-scale Evaluation.

Preliminary Issues, Objectives, and Activities

Site management's first major decision necessarily focused on whether

or not the PCRS was relevant to solving two problems they were most concerned about: (1) low productivity, and (2) low morale, especially as it related to high turnover and high number of grievances filed. These problems, however, can't always be helped by PCRS. For example, if they are primarily due to the combined impact of workforce distrust of management, coupled with chronically bad union/management relations, then attempts to implement the PCRS could make the problems worse. This becomes very apparent when the PCRS requirements involving performance measurements, establishing work standards, and method of payment of incentive bonuses are considered.

After site management had decided to implement the PCRS on a trial basis, an orientation procedure was carefully prepared to inform the workers, supervisors, and union in an accurate and meaningful way of the implications of implementing the PCRS. The importance of this step cannot be overemphasized. Without such an orientation, it is highly unlikely that the broad-based support and cooperation that are indispensable for successful implementation would have been received.

PCRS Development, Administration, and Testing

Most of the important steps required for the initial development, administration, and testing of the PCRS are summarized below:

1. Questionnaire Development. This required a careful definition of the problems involving productivity, morale, and related issues, and discussions of how the PCRS might affect them. Information for this purpose was collected from many sources, including managers, union representatives, workers, supervisors, staff specialists, and others. The information provided the basis for developing a prototypic questionnaire dealing with, among other things, the general conditions of work, the way employees viewed their capabilities to attain various levels of performance, employees' expectations of rewards associated with such levels, and the importance of those rewards to recipients. This information was invaluable for the PCRS development because the information partially answered two fundamental questions: (1) Did the data transcribers perceive themselves as capable of increasing their performance if effective incentives were available?, and (2) What types of rewards would be effective as performance incentives?

2. Workflow Analysis, Performance Measurement, and Results Feedback.

The PCRS requires an analysis of the work process, and elimination of inefficiencies if possible. At the test site, for example, each data transcriber formerly picked up her work at the supervisor's station, and returned the work there upon completion. This necessarily reduced the time that each data transcriber could spend operating the appropriate machines. These procedures were modified so that each supervisor passed out and collected the assigned work, which permitted the data transcribers to be more productive through having fewer disruptions of machine-operation time. After all the modifications were completed, means were devised to measure accurately the performance under the new work procedures, and to feed back the results quickly and regularly. A weekly Operator Analysis Reporting System (OARS) was developed for this purpose.

3. Goal-setting, Criterion Development, and Work Standards Derivation.

The objectives of management and the capabilities of the performance reporting system (OARS) described in the previous section were the primary bases used for establishing performance goals and designating appropriate criteria for the Management Information System Department in which the data transcribers were located. In addition, performance standards were developed for the 190 different procedures performed by data transcribers in this directorate. Collectively, these steps are extremely important to the long-run success of the PCRS. Unless they are done well, the PCRS will inevitably fail--regardless of the managerial support and workforce cooperation afforded it.

After the performance standards had been derived, it was possible to compare individual performance with the standards to determine the relative efficiency of each data transcriber. The degree that a specific data transcriber's efficiency exceeded the standards for assigned tasks was the sole basis for determining the size of financial bonus that the individual received. Thus, incentive awards were strictly contingent on performance--nothing else.

A common complaint of managers regarding some types of incentive plans is that they require too much of their time to administer the plans equitably and smoothly. In contrast, a significant attribute of the PCRS described here is that it requires little management decision-making once it is fully implemented. This is because records of performance, bonuses accrued, bonuses paid, etc. are all essentially accomplished by the OARS report. Thus, required managerial guidance of the PCRS after it is fully operational is minimal.

4. Bonus-Payment Procedure. The amount of bonus earned is calculated on a weekly basis and accumulated until a minimum of $25 is reached, at which time the individual could request payment. Bonuses are paid monthly via checks separate from regular payroll checks. Separate checks are issued because the bonus checks are not drawn from "Compensation" funds, but from funds administered by the Incentive Awards Division of the test site. In addition, separate checks more clearly identify the bonuses as representing superior productivity; such identification may help motivate each individual toward sustaining that productivity.

5. Supervisory Training. The PCRS implementation required several important changes in the activities and responsibilities of the supervisors involved, especially regarding work distribution.

PCRS Preliminary Evaluation

A preliminary cost/effectiveness evaluation of the PCRS was started after the initial results appeared to stabilize into meaningful patterns. After one full quarter of operation tentative conclusions about the degree of PCRS effectiveness could already be made, as described below.

PCRS Modification and Maintenance

Even before the preliminary cost/effectiveness evaluation was begun, some necessary modifications were identified. They included, for example, giving special training to the supervisor in charge of the Digital Computer

Operations Branch, who was responsible for, among other things, dealing with PCRS issues that affected different supervisors on various production shifts. Similarly, an "Incentive Management Coordinator" was trained to initiate resolution of PCRS issues affecting various staff functions such as payroll, comptroller, industrial engineering, and industrial relations.

To provide for continuity of the PCRS, a manual will later be developed to provide detailed guidance for updating and generally maintaining the PCRS near its maximum effectiveness. The manual will contain guidance, for example, for detecting when the work standards must be changed due to the tasks having been modified, the technology altered, worker skill-levels changed, or other important developments have occurred in the workplace. The importance of timely modification of the PCRS to adapt to workplace dynamics cannot be overstressed (Fein, 1971).

PCRS Full-Scale Evaluation

After additional required modifications have been incorporated into the PCRS, some of which will be described below in the discussion section, a more comprehensive cost/effectiveness evaluation will be done when the PCRS has had an adequate period of full operation. That evaluation will address the basic question of whether or not the PCRS has effected substantial cost savings while not reducing (1) the level, efficiency, and quality of production, or (2) the long-run effectiveness and quality of the workforce.

This evaluation will be made on a comparative basis across similar sites. Such comparisons can provide valuable information about the cost/effectiveness of the PCRS at a given site relative to other sites which are and are not using it.

If the evaluation gives convincing evidence that the PCRS is desirable in the department used for the original test, management may also wish to test the PCRS in other departments that are different in tasks, workers, and setting from the original test situation.

Analysis

Rationale and Scope

The analysis in this report is conducted in accordance with the fixed-effectiveness mode of cost/effectiveness analysis (Fabrycky & Thuesen, 1974; Kazanowski, 1968). This requires that the preferred alternative be chosen on the basis of cost, given that all alternatives are equivalent in effectiveness. The two alternatives in the present case are, of course, the PCRS versus non-PCRS production conditions at the test site.

This analysis was strictly limited to test-site impact. This was very important in determining which costs were to be included in the analysis. The research team's salaries and related costs, for example, were not included due to being outside the scope specified.

Field-Test Design

The design for this field test of the PCRS was essentially a before-

and-after contrast that used identical subjects and no control group. There were several important reasons why no control group was used. First, replication data were available from a similar site also implementing the PCRS. Second, and more importantly, use of a control group might result in adverse effects on the long-run motivation and morale of those subjects designated at the test site as being ineligible to receive economic incentive payments during the trial period.

Selection of the base period against which the trial-period results could be contrasted involved two primary considerations--length of period and appropriate calendar interval. In terms of length, one full quarter (13 weeks) was estimated to be long enough for trial-period results to demonstrate the PCRS impact sufficiently to permit a meaningful preliminary evaluation.

Choosing the calendar interval for the base period was more difficult. Before-and-after periods generally should be contiguous in time for the most appropriate contrast. In the case at hand, however, there was concern that possible awareness by the subjects of preimplementation discussions among site management could have influenced the base-period data in some unknown way. Such influence could undermine the credibility of the before-and-after contrast. To preclude that from happening, the calendar interval representing the base period was selected as beginning approximately 6 months prior to the beginning of the trial period. Thus, the base period extended from 5 July through 2 October 1976; the trial period extended from 17 January through 16 April 1977.

Technique

The fixed-effectiveness mode of cost/effectiveness analysis was applied to the present case as follows:

1. All costs incurred by the PCRS during the trial period were determined. These costs were then split into nonrecurring costs associated with setting up the PCRS and recurring costs associated with actual production. Production costs associated with the PCRS were divided by keystroke output during the trial period to provide its cost-per-keystroke. Similarly, production costs of the base period were divided by its keystroke output to provide the cost-per-keystroke for that period.

2. The difference in cost-per-keystroke between the periods was multiplied by the keystroke output of the trial period. This provided the production-cost savings associated with PCRS implementation.

3. An evaluation was made to determine if implementing the PCRS had diminished production "effectiveness" while the production-cost savings were being generated.

4. Net savings for the trial period were determined by subtracting set-up costs from production-cost savings. Savings projections based solely on trial-period parameters were derived.

5. Savings projections based on specified levels of aggregation and generalizability were derived.

RESULTS AND DISCUSSION

PCRS Set-Up Costs

The nonrecurring costs incurred during setting up the PCRS at the test site are detailed in Appendix C and summarized below:

1. Recorded Costs

a. Equipment Purchased	$ 855.55
b. Software Development	3693.00
c. Personnel Training	3562.18
Total Recorded Cost	$8110.73

2. Possible Unrecorded Costs

(Estimate 10% of total Recoded Costs)	811.07
Total Set-up Costs	$8921.80

The above figures represent every reasonable effort to include all possible nonrecurring set-up costs associated with implementing the PCRS. Thus, in addition to the explicit costs (e.g., equipment purchased), there was also inclusion of implicit cost such as (1) decreased keystroke production while data transcribers and supervisors were being trained during work hours for PCRS operations, and (2) use of test-site staff personnel, such as the computer programmer who developed the required software. Finally, a substantial adjustment for possible nonrecorded costs was included.

Cost Savings

The overall cost savings of the PCRS relative to costs of former production conditions at the test site can be informatively illustrated by focusing separately on (1) savings generated by production-cost reduction, and (2) the net savings remaining after all set-up costs were absorbed.

Production-cost Savings

Deriving the production-cost savings required comparative data on two basic dimensions--production costs and production output. Production costs of the data-transcribing activities at the test site were primarily accounted for by the Recharge Rate (RR) already described in this report under Measures. It should be recalled here that the RR is a composite cost, derived and periodically updated by the test-site comptroller, that represents the current overall cost of the data-transcribing activity on an hourly basis. Component cost levels of the RR, detailed in Appendix A, document cost reflecting the following aspects of the data-transcribing operation: (1) basic salary of data transcribers, (2) government portion of data transcriber's pension and other benefits, (3) machines used, (4) supervisor salaries, and (5) overhead.

Production costs for the base period and trial period are compared in Table 1. Comparative production output, in terms of cumulative keystrokes,

for the base period and the trial period were 35,554,496 and 37,117,213, respectively. These figures included the combined print/verify activities of all shifts involved in the data-transcribing operation at the test site, as detailed in Appendix D.

Given the comparative production cost and production output presented above, the production-cost savings can be derived by multiplying the inter-period difference of cost-per-keystroke by the keystroke output of the trial period. These relationships are illustrated in the following formula:

$$\left[\frac{\text{Production cost (base)}}{\text{Production output (base)}} - \frac{\text{Production Cost (trial)}}{\text{Production output (trial)}}\right] \text{Production output (trial)} = \text{Production-cost savings of trial period}$$

Inserting appropriate values into the formula gives the following results:

$$\left[\frac{\$84,920.43}{35,554,496} - \frac{\$78,370.03}{37,117,213}\right] (37,117,213) = \$10,281.47$$

The results indicate that $10,281.47 in production-cost savings were generated during the 13-week period that the PCRS was evaluated on 17 fully-qualified civil-service data transcribers.

Net Savings

The net savings generated during the trial period were the production-cost savings remaining after subtracting all set-up costs. Thus, the production-cost savings of $10,281.47, when reduced by the total set-up costs of $8,921.80, leave a net savings of $1,359.67.

Net savings of the PCRS trial period, as derived above, must be carefully interpreted. The primary issue revolves around whether or not it is appropriate to absorb during the trial period itself all the nonrecurring set-up costs incurred during the PCRS implementation. An alternative way to recover the set-up costs is, of course, to distribute a pro-rated portion of them over a specified number of accounting periods. This alternative increases the net savings for the trial period, but reduces the production-cost savings during the periods over which the set-up costs are distributed. In this study, however, the set-up costs were totally absorbed in the trial period for the two following reasons.

First of all, completely absorbing during the trial period all the PCRS set-up costs demonstrated emphatically that such costs were less than the savings generated solely from reductions in production costs during the same period. This is very important information. It clearly shows that nonrecurring set-up costs were fully recoverable in less than one full quarter of operation. Few investments have such short "payback" periods.

Furthermore, the probable recovery periods for other Navy sites implementing the PCRS would be even shorter when consideration is given to

economies of scale that are possible. That is, the set-up costs of similar Navy sites implementing the PCRS can reasonably be expected to decrease substantially. Such decreases could occur, for example, by appropriately adapting the testsite-developed software to meet the relevant particulars of other Navy sites. These adaptations could effect substantial reductions in set-up costs when contrasted with the alternative of each site developing its software "from scratch." Similarly, the PCRS-oriented questionnaire that was developed at the test site presumably could be readily adapted to other sites, and thereby effect another major reduction in set-up costs.

The second primary reason for absorbing all the nonrecurring set-up costs during the trial period itself was that it facilitated interpretation of the cost-savings projections, presented below, that were based soley on test-site parameters. In particular, this manner of handling the set-up costs permitted the cumulative savings value for the specified outyears to be derived as the sum of (1) the trial-period net savings (i.e., after all nonrecurring set-up costs recovered) and (2) post-trial period savings based solely on production-cost reduction (i.e., without adjustments for distributed set-up costs).

Interrelationships among the PCRS set-up costs, net savings, and production-cost savings have now been evaluated in terms of impact on the monetary results of the trial period. It was demonstrated that , based purely on reduction of production costs, savings were found to be in excess of $10,000 for the PCRS trial period. Given that the trial period involved only 17 subjects for 13 weeks, such savings are noteworthy--especially when generated during a period when the inevitable problems associated with implementing any new major system had to be solved. What still remains to be evaluated, however, is the impact of PCRS implementation on the non-monetary dimensions of production "effectiveness."

PCRS Impact on Production Effectiveness

The fixed-effectiveness mode of cost/effectiveness analysis permits competing alternatives to be selected solely on the basis of cost only if all alternatives are equivalent in effectiveness (Fabrycky & Thuesen, 1974; Kazanowski, 1968). Superiority of the PCRS relative to former production conditions at the test site has already been demonstrated in terms of costs. Not yet demonstrated, however, is the impact that the PCRS implementation had on test-site production "effectiveness."

The nonmonetary dimensions of production effectiveness on which the impact of PCRS implementaion was evaluated were (1) level of production, as measured by total keystrokes, (2) efficiency of production, as measured by keystrokes per man-hour, and (3) quality of the production process, as measured by (a) number of higher-cost overtime hours used to keep workload backlog within acceptable limits, and (b) number of average daily batches (defined previously under Measures) in workload backlog. Quality of production output--as measured by the keystroke error rate--was determined by site management to be within acceptable limits even before the PCRS was implemented. Therefore, quality of output was not explicitly evaluated during the PCRS field test. It should be noted, however, that since the quantity of output is strictly limited to keystrokes "verified" as correct, the quality of output is essentially held constant across the inter-period comparison of

overall production effectiveness. Results of the production effectiveness evaluation are summarized in Table 2.

The results in Table 2 clearly indicate that the overall production effectiveness of the test site was not diminished by PCRS implementation when evaluated in terms of the nonmonetary units specified. Moreover, each of the respective values of data corresponding to the dimensions of level, efficiency, and quality of the production process were better during the trial period than the base period. Tests for statistical significance are not required, therefore, because any statistically significant differences found could only demonstrate the superiority of production effectiveness during the trial period. In terms of the fixed-effectiveness mode of cost/effectiveness analysis, however, it is only required to demonstrate that the alternative preferred on the basis of cost (i.e., the PCRS) is not inferior on basis of effectiveness (Fabrycky & Thuesen, 1974). Results in Table 2 clearly show this to be the case. In addition, these results from preliminary cost/effectiveness analysis may have understated considerably the full potential of the PCRS for long-run cost reductions of data-transcribing activities. Some of the major reasons for the probable understatement are the following:

1. The increase in productivity during the trial period led to elimination of the workload backlog accumulated prior to PCRS implementation; consequently, there was often inadequate work available to keep all data transcribers busy. These conditions necessarily lowered the potential productivity increase which could be demonstrated by the PCRS during the trial period.

2. Many of the important ad hoc problems that inevitably accompany implementation of a major system such as the PCRS had to be solved during the trial period; this decreased the productivity during this period relative to test-site productivity after the trial period ended.

3. Raw data from a site similar to the test site appear to be superior--or at least, equal--to data from the test site for comparable stages of the trial period.

4. The research team could detect no informal evidence toward the end of the trial period and thereafter that suggested the PCRS was having other than a primarily positive influence on workforce variable such as turnover, absenteeism, supervisory relations, union/management relations, and morale. However, temporary test-site limitations in information sources relating to these variables necessitated their exclusion from the preliminary cost/effectiveness analysis. Thus, the results presented in this study probably undervalue the long-run cost savings attributable to the PCRS impact on data-transcribing activities at the test site.

PCRS impact on production costs and the level, efficiency, and quality of the workflow has been tested--but the impact on the workforce itself remains unknown. Because the PCRS impact on the workforce dimensions listed above has important implications for the long-run trends of the workflow, it is necessary that those dimensions be systematically evaluated. To provide such information requires that an appropriate information system will have to be developed and implemented at the test site. As a minimum, the system must provide information that will permit accurate answers

to the following questions: (1) What are the production, administrative, and other organizational effects that are associated with major changes in the level of a workforce variable (e.g., turnover)? and (2) How can such effects be quantified and converted into appropriate monetary units that will enable cost/effectiveness analysis?

Developing and implementing the information system described is a complex undertaking. Moreover, the literature on human resource accounting is still evolving in terms of the theoretical relationships on which such information systems are based, and results from field tests of such relationships have not established a comprehensive and totally consistent pattern of findings (Flamholtz, 1974; Likert, 1967). As a result, both the theoretical and empirical guidance regarding such information systems is limited. Developing and successfully implementing the information system will therefore require successive modifications of the system until satisfactory results in terms of accuracy and cost of the information provided are achieved.

If an information system can be implemented at the test site that will satisfactorily provide the information required to include workforce variables in a cost/effectiveness analysis of the PCRS, the same system will also be implemented at another Navy site already having the PCRS in operation. Then, in contrast to the present analysis which was strictly limited to production costs and workflow variables, a more comprehensive cost/effectiveness analysis can be conducted that will take into account the following : (1) multiple sites, (2) a much larger number of data transcribers, (3) a longer trial period, and (4) multiple cost criteria that reflect workforce as well as workflow variables. Overall, results from an analysis expanded in this form would add greatly to understanding the true cost/effectiveness impact of the PCRS on data-transcribing activities in Navy sites. Until such a comprehensive analysis is done, it must be concluded that results from the preliminary analysis reported in this study indicate that the PCRS appears to have considerable potential, but awaits broad-based confirmation from further field tests.

Test-Site Savings Projections

The net savings associated with implementing the PCRS were generated from 17 data transcribers during a 13-week period. While those savings were noteworthy in and of themselves, of far greater interest to test-site management and others is the cumulative value of the PCRS savings when projected through specified outyears. Projections based on a savings rate identical to the trial period and representing 1, 3, and 5 years are shown in Table 3.

The projections list cumulative values via combining the actual net-savings generated during the trial period and the projected production-cost savings generated after the trial period. The production-cost savings are compounded monthly because the test-site data transcribers receive cash incentive bonuses on a monthly basis if the data transcribers are eligible and request payment--in contrast to letting the bonuses continue to accumulate. Since such bonuses represent a portion of the production-cost savings that have already accrued at the test site during the monthly accounting period, projections based on coinciding compounding and periodic-payment intervals are warranted (Fabrycky & Thuesen, 1974). In the context of the present case, this simply means that the production-cost savings previously accrued are compounded on the same date that the next monthly increment of savings is accrued.

The lump-sum savings are also compounded monthly in order to coincide with the compounding cycle of the production-cost savings. This facilitates interpretation of the cumulative projected value derived from combining the actual trial-period net savings and the projected production-cost savings. It should be noted, however, that monthly compounding of the lump-sum savings probably represents a conservative bias in the overall projections. This is based on the fact that many financial institutions would compound such lump-sum savings on a "continuous" basis, which generates a higher yield. For details on how the compounding factors were derived and related issues, see Appendix E.

Navy Community Projections

While projections of PCRS savings based solely on trial-period parameters provide valuable information to test-site management, such projections are of limited utility to the Navy community as a whole. To be meaningful from that perspective, the projections must reflect (1) specified levels of aggregation of civil-service data transcribers in other Navy organizations, and (2) specified levels of generalizability of test-site results to other Navy sites. From this point forward, therefore, this report will necessarily deal with issues stemming primarily from test-site results, but which have important implications for other sites and activities of the Navy community in which the PCRS may be implemented. Some of these implications will be addressed in sections below that are respectively designated for discussion, conclusions, and recommendations.

Levels of Aggregation

To derive projections of PCRS cost savings that represent progressively wider scope, civil-service data transcribers in Navy organizations were aggregated at three overlapping levels: (1) the full complement available at the test site, (2) the approximate number available in the shipyard community, and (3) the approximate number available in the NAVMAT community.

There were 26 civilians performing data-transcribing activities at the test-site shipyard during the PCRS trial period, some of whom were not eligible for inclusion in the PCRS evaluation as reported in this study. Though all data transcribers at the test site were operating under PCRS conditions during the trial period, the criteria for selecting data transcribers for the PCRS evaluation were described previously under Sample.

At the level of the shipyard community, a survey of personnel staffing levels of the individual shipyards located at least 200 civil-service data transcribers. At a still higher level of aggregation, the same survey located at least 725 data transcribers in shipyards, supply centers, ordnance depots, and other activities in the NAVMAT community that require extensive data processing.

Levels of Generalizability

The broad issue of how well the test-site results can be generalized to other Navy sites can be split into subordinate issues dealing primarily with (1) credibility of the trial-period results as being representative of the long-run results at the test site, and (2) comparability of other Navy

sites to the test site. Splitting the issues in this manner is somewhat artificial in that many of the issues are heavily interrelated. The splitting is helpful, however, in illustrating that generalizability involves two primary components which, though interrelated, can be meaningfully evaluated separately. After the credibility issues and comparability issues are qualitatively evaluated, their combined impact on the cost-savings projections will be quantitatively derived.

Credibility

Evaluating the following issues will assist in determining how credible the trial-period results are in terms of being representative of the long-run results at the test site if the PCRS is continued in operation. Or, in terms of "internal validity" (Cook & Campbell, 1976), to what degree can the trial-period results be attributed to the PCRS versus effects from other causes?

Field-Test Design and Trial-Period Parameters. In essence, the PCRS field test consisted of comparing data across two 13-week periods on the same 17 subjects with no control group; the primary reasons for using no control group were explained previously under Analysis. Thus, given the nature of the field test and the limited scope of this preliminary cost/effectiveness analysis, it is not appropriate to assert that the results are unequivocably due to the PCRS. The cost savings derived from higher productivity could, for example, stem mostly from what is broadly known as the "Hawthorne Effect." If true, this means, among other things, that the results are not primarily due to the inherent features of the PCRS, but due simply to the subjects' reaction to the heightened attention that management and the research team focused on the subjects' individual and collective productivity during the trial period. The increased productivity during the trial period may not, therefore, be representative of a long-run trend. (It is important to note, however, that the research team had been actively working with the subjects for over a year before the PCRS came into effect--during which time no increase in productivity was observed.) Alternatively, it is far more plausible that the data transcribers were motivated toward higher productivity by the performance-contingent economic incentives. If true, the increased productivity should be maintained as long as the PCRS remains in effect. Whatever the case, the question of what was the primary determinant during the trial period of the increased productivity, and the implied duration of such productivity, cannot be answered conclusively from results of the field test and preliminary analysis described in this report; this becomes increasingly evident as the following issues are evaluated.

It should be noted, however, that this research has not been concluded. Data from multiple sites, additional performance measures, and more comprehensive and in-depth analyses will be used in later evaluations.

Impact of Critical Incidents and Extraneous Factors. Shortly before the trial period began, the director of the Management Information System Department at the test site resigned. The true impact of this event is difficult to measure because there is no control group. It is very possible, however, that the event depressed the possible increase in productivity associated with the PCRS during the trial period. This is because the

data transcribers may have experienced uncertainty regarding whether the new director would continue the incentive program. If discontinued, their increased productivity would be unrewarded. In overall effect, therefore, the change of directors may have depressed the cost savings generated during the trial period, despite the fact that the change occurred after the base period had ended and before trial period had begun. This is because the cost savings associated with the PCRS were measured, in part, as the difference in production costs between the base period and the trial period, and the incident may have had a deferred impact on the latter period.

Another incident that could have impacted on the results involved exchanging the shift supervisors on the day and swing shifts between the base period and the trial period. The incident had important implications because the day shift had a considerably larger number of data transcribers assigned to it than the swing shift. Thus, if the increased productivity during the trial period was primarily due to supervisory practices, per se, and not due to the PCRS, then the exchange of supervisors could have had a major impact on the results. As disclosed in Appendix D, however, keystroke productivity for the swing shift increased proportionally more than the day shift when production efficiency in terms of keystrokes per man-hour is contrasted within shifts across periods.

The extraneous factor that may have had the greatest unmeasurable impact of all on the trial-period results stems from the fact that the data transcribers frequently ran out of work during the latter stages of this period--even the large workload backlog accumulated before the PCRS was implemented was eliminated during this time. If the increase in productivity during the trial period was, in fact, due to implementation of the PCRS, then this extraneous factor is significant for two reasons: (1) it imposed a situational constraint on the amount of potential productivity increase that the PCRS could demonstrate on the basis of a 13-week trial period, and (2) it may have seriously undermined the data transcribers' motivation toward sustaining the increased productivity due to their possible anxieties over managements' potential reactions to the data-transcribing operation being chronically "overmanned," which primarily resulted from increased productivity during the trial period. Such possible anxieties have their foundation in the fact that management issued no policy statement before or during the trial period to the effect that any overmanning caused by increased productivity would be remedied by normal attrition, voluntary transfers, etc.--as opposed to reductions-in-force, involuntary transfers to other departments, and similar undesirable alternatives.

There is a strong possibility, therefore, that the inadequate workload and its possible consequences restrained the PCRS during the trial period from demonstrating its full potential for increasing productivity.

Motivational Durability of the PCRS. Another important factor in determining how credible the trial-period results are in terms of being representative of the long-run results at the test site if the PCRS is continued in operation is, of course, the inherent features of the PCRS itself. Of particular significance are those features involving (1) the derivation, composition, and level of work standards, (2) the level of

economic incentives to be awarded for productivity exceeding work standards, and (5) the computation and timing of productivity-bonus payments. As described previously under Procedure, do these features of the PCRS collectively form a functional and administrative foundation capable of "motivating" data transcribers toward higher productivity through the long run?

Comparability

In addition to how credible the trial-period results are in terms of the long-run results at the test site if the PCRS remains in operation, the generalizability of the test-site results to other Navy sites is greatly dependent on the overall comparability between the test site and other sites. Or, in terms of "external validity" (Cook & Campbell, 1976), to what degree can trial-period results from the test site be generalized to other persons, times, or settings? In particular, to what degree can the test-site results be replicated at other sites in the shipyard community or NAVMAT activities that implement the PCRS in data-processing departments?

Some dimensions that management of other sites considering the PCRS implementation should evaluate are addressed below:

1. Similarity of Basic Technology and Workforce to That at the Test Site. That is, how comparable are the data-processing equipment, work tasks, and quality of data transcribers at other sites to the test site?

2. Site History. If recent history includes failure of a previous economic incentive program, support and cooperation from data transcribers and other affected personnel cannot reasonably be expected to be high during implementation of a similar program such as the PCRS.

3. Workforce "Trust" in Management. If the workforce believes that increased productivity during the PCRS trial period might be used by management to set higher standards for "normal" productivity later, then the trial period will not likely demonstrate significant productivity increases. Similarly, if the workforce believes that increased productivity resulting from PCRS implementation might ultimately lead to reductions-in-force due to overmanning (assuming workload remains constant), then--again--the PCRS trial period will likely not result in noteworthy productivity increases. A policy statement by top management that would address both these issues would benefit any site considering PCRS implementation.

4. Union Support. If the unions involved do not support the PCRS on most of its basic features, then its ultimate failure is almost a certainty.

This concludes the qualitative evaluation of issues relating to credibility of trial-period results and the comparability of the test site to other Navy sites. In summary, it should be noted that the test-site results apparently can be substantially generalized to other Navy sites. This tentative assertion is based on preliminary data from another Navy

shipyard in which the PCRS has also been implemented on a trial basis. Although results from this site have not yet been summarized in a manner permitting statistical comparisons with the test site, the raw data appear to be as good or better than those of the test site for comparable stages of the trial period.

Quantifying the Credibility/Comparability Issues

The interdependent nature of the credibility and comparability issues can be quantified to estimate their combined impact on the generalizability of test-site PCRS savings to other Navy sites. This is done by assigning numerical weights to the trial-period results in accordance with their subjective value in terms of credibility and comparability implications. For example, if after careful consideration of test-site conditions and events during the trial period, a manager at a site contemplating PCRS implementation decides that the trial-period results described above were probably lower than what they will generally be for equal-length periods at that test site through the long run, the manager would assign an appropriate numerical weight greater than one (e.g., 1.25) to compensate for the understated trial-period results. This numerical weight would then represent the collective credibility issues in determing the generalizability factor.

Similarly, if the same manager identified important conditions at his site that definitely diminished the comparability between the test site and his own, he would assign an appropriate numerical weight less than one (e.g., 0.75) to compensate for the incomparabilities noted. This numerical weight would then represent the collective comparability issues in determining the generalizability factor.

The separate numerical weights are then multiplied together to form the product that will represent the overall generalizability of the PCRS cost savings generated during the test-site trial period to the particular conditions at his own site. Thus, the generalizability factor if the numerical weights in the example are used is (1.25)(0.75) = .9375. Given that the production-cost savings associated with the PCRS during the test-site trial period were $10,281.47, then ($10,281.47)(.94) would be an appropriate estimate by the manager of the cost savings that the PCRS would generate during a trial period of equal length at his own site if the same number of data transcribers (17) were used.

Projections Reflecting Combined Aggregation/Generalizability Impact

Projections of production-cost savings that reflect the combined impact of levels of aggregation and levels of generalizability are shown in Table 4. The projections are not based on net savings, i.e., after set-up costs are absorbed, for the following reasons: (1) PCRS set-up costs are so small relative to production-cost savings that such costs were recoverable in less than one full quarter of operation at the test site; therefore, set-up costs have minimal significance in long-run projections of PCRS savings, and (2) set-up costs at the test site are amenable to economies of scale from the perspective of other Navy sites. That is, several end-products generated from set-up costs at the test site--such as the software package and PCRS

questionnaire, for example--can be readily adapted to the data-transcribing activities at other Navy sites.

Table 4 is self-explanatory. Special note should be given, however, to the extensive range of the projected-savings values that correspond to different levels of aggregation and generalizability. Within the 3-year projections, for example, while at the generalizability level of 1.00, the value of the production-cost savings ranges from $221,000 to $1,700,000 to $6,200,000 when aggregated from the "Test-site Shipyard" to "All Shipyards" to the "NAVMAT Community." Similarly, within the 3-year projections and remaining at level of aggregation of All Shipyards, the value of savings ranges from $849,000 at the 0.50 level of generalizability to $2,500,000 at the 1.50 level of generalizability.

Due to this extensive range in value of the projected savings, it is important that managers having control over the various aggregation levels carefully evaluate the credibility/comparability issues that underlie the generalizability of the test-site results to other Navy sites. Without such evaluation, expectations of savings to be derived from PCRS implementation at other Navy sites may be extremely inaccurate.

Integration of Results, Projections, and State of the Art

Early in this report it was documented that the state of the art pertaining to the general effectiveness of economic work incentives was highly controversial and, moreover, very contingent on the overall work situation in which they were applied. PCRS results described in this report, however, represent data from only 17 subjects at a single test site for a trial period of 13 weeks. Yet, projections based on those results were made for periods extending through 5 years and at levels of aggregation that reflect hundreds of data transcribers from diverse Navy sites. This raises a strongly implied question: Are such projections based on the semblance or substance of realism? If based on the latter, what inherent features of the PCRS presumably warrant such long-term and broad-based projections. In relation to data-transcribing activities, some of those features are the following:

PCRS Incentive Rewards Are Tightly Linked to Documented Performance--Nothing Else

As a result of being strictly contingent on demonstrated performance, the motivating value of the PCRS for increasing productivity should not diminish as a result of gradually developed expectations that the incentive rewards will occur as a matter of course. Incentive programs can lose their effectiveness when the tight linkage between performance and rewards becomes loosened through, for example, rewards being based on inaccurate evaluation of performance by supervisors using subjective bases.

Data-transcribing Activities Provide a Precise Measure of Performance

Few measures of performance can be more objective and precise than keystroke per operator per time period. The PCRS is not hampered, therefore, by a technology that doesn't provide a precise measure of the performance on which the incentives are contingent.

Each Data Transcriber Essentially Has Full Control Over Self-Performance

In data-transcribing activities, there is very little performance interdependence with other members of the work group. Each worker's performance is therefore primarily a function of individual motivation, capability, and little else. In contrast, some incentive plans focusing on the individual worker fail because the workers perceive too much performance interdependence with co-workers; this causes the individual efforts, outputs, and rewards to be inadequately differentiated across the various workers. As a result, the higher-performance workers may become de-motivated because individual rewards do not correspond closely to individual performance.

Data-transcribing Activities Permit Short Feedback Cycles

Machines on which the data transcribing is done directly record the number of keystrokes for each operator. Under PCRS, these performance totals are accumulated weekly and corresponding bonus payments are paid monthly. Incentive plans can become gradually ineffective if the workers feel they are not being frequently and adequately informed regarding the status of their performance and implied bonus payments.

Economic Incentives Have High Utility For Most Data Transcribers

Pay grades of data transcribers are generally restricted to GS-3 or GS-4 levels. The economic incentives of the PCRS can therefore be reasonably expected to have high utility for the heavy majority of data transcribers. Incentive plans can fail because the incentives used--whether economic or otherwise--have low utility for the workers involved, with the result that high performance is not elicited.

PCRS Set-up Costs for Data-transcribing Activities are Relatively Low

Such costs were fully recoverable from production-cost savings in less than one full quarter of operation. Thus, Navy managers need not be deterred from testing the PCRS because of initial cost-outlay required. Moreover, economies of scale are possible because several types of end-products (e.g., software development and questionnaire development) generated from set-up costs at the test site can be readily adapted to other Navy sites. This differs from incentive plans that are sometimes not given a fair field test because the required "sunk" costs would be prohibitive if the test failed.

PCRS Administrative Costs are Relatively Low for Data-transcribing Activities

Once the PCRS is completely implemented and fully operational, the recurring administrative costs for special record-keeping, computer, payroll, and miscellaneous services are small relative to the production-cost savings that are generated. In contrast, it could be necessary to terminate an incentive plan after it becomes fully operational if the time, effort, and other resources required from supervisors, managers, industrial engineers, etc., remain so costly that they negate the benefits of productivity increases from production workers.

In summary, the above features of the PCRS, as related to data-transcribing activities, constitute the primary basis for the long-term projections that were derived from results of a field test of limited scope. One more major issue, however, remains to be addressed: Will the present data-transcribing activities in the Navy community undergo such massive technological and other changes that projections based on the field-test results will soon become largely irrelevant, regardless of their previous credibility? This question is extremely important, and equally difficult to answer. Representatives of the test site's Management Information System Department, however, foresee no massive changes in data-processing workload, machine technology, and data-transcribing tasks that will occur soon enough to invalidate the 5-year projections. (Apparently the logical event for precipitating such possible changes would be a breakthrough in optical character-recognition technology.)

Less massive changes in the data-transcribing activities can, of course, be readily accommodated by the PCRS through updating the work standards, revising the incentive-bonus rate, and similar adjustments. In conclusion, therefore, the PCRS relevance to data-transcribing activities remains undiminished for the foreseeable future, based on informal judgments of practitioners.

Organizational Implications

Major implications from implementing the PCRS extend far beyond the data-transcribing activities to which the PCRS is directly applied. Several broad organizational implications which management of each site contemplating PCRS implementation should evaluate are the following:

Potential Perceptions of Wage and Salary Misalignment

The implications of perceived misalignment can be best illustrated by addressing the vertical and horizontal perspectives of this issue separately:

Vertical Perspective. The vertical perspective primarily involves the impact on the relative earnings of the supervisors of the PCRS data transcribers. If the average earnings of the data transcribers increase and those of the supervisors don't, the earnings differential will be narrowed. As a result, the supervisors responsible for the group performance of the data transcribers may lose motivation to fully support the PCRS because they would earn less relative to their subordinates than before the PCRS was implemented.

A potential remedy to this problem is to give some portion of the average monetary bonus of the data transcribers to the supervisors. The additional cost that this implies would best be borne by management because the current bonus at the test site is only 11 percent of the cost savings derived from each data transcriber's increase in productivity. Thus, 89 percent of the savings redound to management, which provides ample basis for redistributing some of those savings to supervisors who are instrumental in producing the cost savings through better work distribution and other supervisory practices required by the PCRS. Alternatively, if the supervisory bonus-sharing is deducted from the data transcriber's portion of

the production-cost savings, their motivation toward maintaining the higher productivity may be seriously undermined because the bonuses under these conditions may be too small to have incentive properties. Moreover, it should be noted that the current 11 percent portion of the cost savings is already an extremely low sharing rate relative to incentive plans in the private sector (Fein, 1971).

Test-site management and the research team are already evaluating alternatives that would give the supervisors a vested economic interest in the collective performance of data transcribers operating under PCRS conditions.

Horizontal Perspective. Of far greater difficulty is devising a remedy for correcting the perceived horizontal misalignment--i.e., perceptions based on the perspective of non-PCRS employees who do not have a hierarchical relationship with the PCRS employees. The perceived misalignment is especially intense if the non-PCRS and PCRS employees work in close spatial proximity or have a highly interdependent workflow. From a broader organizational perspective this perceived misalignment is very important because the PCRS may have a strongly beneficial effect on the motivation, performance, and morale of the data transcribers while non-PCRS workers in the same organization may simultaneously experience a strongly detrimental effect on the same dimensions. In great part, the detrimental effect on the non-PCRS workers would be based on the following reasons: (1) they are denied opportunity to share in the PCRS economic rewards, and (2) they feel that the data transcribers are deriving economic rewards not based entirely on their own activities, efforts, and performance.

Two possible solutions to this problem are the following: (1) Include the non-PCRS employees under the PCRS if their tasks and other work relationships are amenable to PCRS implementation, or (2) adapt the PCRS from an individual incentive plan to a group incentive plan, on the rationale that many non-PCRS employees have supporting roles that are broadly interdependent with the workflow of PCRS data transcribers. The implications of changing an incentive plan from an individual to a group basis are very complex, however, and must be considered carefully on the many tradeoffs involved (Belcher, 1974).

Conflict Resolution

Implementing the PCRS may precipitate temporary conflicts among various types of employees and organizational functions. The process of setting work standards, for example, may engender conflict among the industrial engineers, union representatives, and supervisors. Similarly, the whole concept of economic incentives being necessary to increase productivity may be challenged by the comptroller. Moreover, the new methods of work distribution involving data-transcribing activities may not at first be accepted wholeheartedly by the supervisors. It is imperative, therefore, that management attempt to exercise a firm and equitable influence toward resolving the conflicts to the acceptance of all concerned during the implementation process.

Management must, to be sure, continue to provide broad-based support of the PCRS after it is fully implemented and operating smoothly. During this period, it is especially important that management give immediate attention to resolving any emerging conflicts involving the PCRS. This could include, for example, handling PCRS-related grievances in a manner permitting quick, fair, and decisive resolution--even if such procedures differ considerably from the normal grievance-resolution process. This is necessary because if resolving serious problems with economic incentive plans is deferred for routine processing, potentially irreparable damage to the work unit's effectiveness, production costs, employee morale, and union/management relations could result (Belcher, 1974; Fein, 1971).

PCRS Discontinuance

It is possible, of course, that management may decide to discontinue the PCRS after it has been implemented on a trial basis at a given site. Whatever the reasons for discontinuance, it may cause extensive damage to the site's workflow if the process of discontinuance is not systematically thought out and carefully executed. It is unrealistic, of course, to expect that if the PCRS is successfully implemented and tested, that it can be discontinued without leaving an aftermath of major proportions in terms of effects on productivity, morale, and union/management relations. In the private sector, for example, managements often use costly "buy out" plans when economic incentive plans are discontinued. Such plans represent an attempt by management to keep peace with unions, to keep productivity at high levels, and to not demoralize the workforce as a result of the decrease in individual earnings that often accompany the discontinuance (Fein, 1971).

To minimize the detrimental impact of the discontinuance on the workforce, the following procedures might be helpful: (1) Management should not make the discontinuance decision unilaterally, or begin its execution without adequate forewarning and explanation being given to all concerned. Reactions of the workers and unions might be more moderate and compliant if they have been adequately consulted during the decision-making process and, hopefully, understand the necessity for the discontinuance as a result, and (2) management should not expect productivity, turnover, absenteeism, and morale to remain at the desirable levels they may have reached while the economic incentive plan was in effect. Management should expect slippage towards pre-incentive levels and, in addition, should be wary of initiating sanctions to prevent this tendency. Such sanctions, if implemented, might be so resented by the workers and unions that open conflict could result.

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

A fixed-effectiveness mode of cost/effectiveness analysis was used to test the following hypotheses:

1. Implementation of the PCRS will substantially reduce the production costs of data-transcribing activities relative to costs associated with former production conditions at the test site.

2. Implementation of the PCRS will not diminish the production effectiveness of data-transcribing activities while the costs are being reduced.

Based on, among other results, the fact that production-cost savings in excess of $10,000 were generated in a field test on 17 data transcribers for 13 weeks--during which time the keystrokes per man-hour increased over 14%--the hypotheses are concluded to be true.

Recommendations

It is recommended that Navy managers having control over sites with large numbers of civil-service data transcribers evaluate the procedures and results of the PCRS field test from the perspective of possible implementation. Such managers should give special attention, of course, to issues underlying the generalizability of the test-site results to sites under their control.

If Navy managers decide that the overall generalizability of test-site results appears sufficient to warrant implementation of the PCRS, it is recommended that they make a firm commitment to provide the resources and broad-based management support required to ensure that the PCRS receives a full and fair field test. Only then will accurate data be available after the trial period to form a reliable basis for determining whether or not the PCRS should be continued.

It is also recommended that appropriate procedures be initiated to evaluate the implications of modifying civil-service policy regardng economic incentive awards (Federal Personnel Manual, Note 2). In particular, the implications should be examined of differentiating the level and bases of economic awards between (1) "one-shot" innovations that result in cost savings due to inventions, new techniques, or similar devices that workers use, and (2) productivity-induced cost savings that stem primarily from additional effort, higher motivation, or superior ability of the workers themselves. The present guidelines seem more appropriate to awards for once-only procedural or "hardware" innovations than to increased individual productivity on a continuing basis.

Current guidelines suggest that workers be awarded no more than 10 percent of the net cost savings, given that the savings are $1000 or less; above that level, it is to be no more than 5 percent of the cost savings

up to $10,000--with additional decrements in the worker's share thereafter. Such rates for worker participation in cost savings may provide adequate motivation for attempting inventions or other nonrecurring innovations that may ultimately lead to massive cost reductions for the site or agency and, in turn, to substantial economic awards for the originators. It is less probable, however, that such low ceilings and decremental sharing rates will provide adequate motivation to keep productivity-induced cost savings at really high levels through the long run. As a comparative benchmark, it should be noted that participation rates for individual-based economic incentive plans in private industry are generally in the 30 to 70 percent range, with no decrements (Belcher, 1974).

The potential significance of this recommendation is that unless the guidelines are revised in a manner that satisfactorily addresses the issues described above, the PCRS may be less than totally effective through the long run--and the ineffectiveness may be due less to inherent features of the PCRS than to current policy guidance regarding the level and bases of economic incentive awards.

The implications of modifying policy on economic incentive awards are, of course, complex and far reaching. One such implication deals with the possible effects on the long-run productivity and morale of those employees not covered by the economic incentive plan, a potential problem for which prospective remedies have already been discussed. This example illustrates that considerable care should be taken to ensure that intricate tradeoffs implied by this recommendation will be systematically evaluated from a broad and long-run perspective.

Table 1

Comparative Production Costs of
Base Period and Trial Period

Costs	Base Period	Trial Period
RR x Man Hours[a]	$83,874.48	$76,760.32
OT x Man Hours[b]	1,045.95	243.21
PCRS Bonus Payments[c]	- - - -	916.50
PCRS Admin Costs[d]	- - - -	450.00
Total	$84.920.43	$78,370.03

[a]Current Recharge Rate (RR) is $10.82 per hour. The Base Period used 7751.8 total man hours; the Trial Period used 7094.3 total man hours. Of the total man hours used in comparative periods, the relative man hours paid at overtime rate are specified in Footnote b.

[b]Overtime rate is 1.5 basic hourly salary. Thus, the $2.02 represents the 0.5 overtime component cost. Overtime man hours for the Base Period and Trial Period are 517.8 and 120.4, respectively.

[c]PCRS bonus-payment calculation is detailed in Appendix B.

[d]PCRS administrative costs were estimated by representative of testsite comptroller. At level of cost listed above, the estimate is intended to cover all possible costs of PCRS requirements for special record-keeping, computer, payroll, and miscellaneous services.

Table 2

Comparative Production Effectiveness of Base Period and Trial Period

	Base Period[a]	Trial Period[b]	Direction & Amount of Change	Direction of Change Desirable?
Level of Production (Keystrokes)[c]	(36,118,853) 35,554,496	37,117,213	+2.76%	Yes
Production Inputs Used[d] (Total Man-hours)	7751.8	7094.3	-8.5%	Yes
Efficiency of Production (Keystrokes per Man-hours)	4587	5232	+14.1%	Yes
Quality of Production				
Excess-Cost Penalties (Overtime Man-hours Used)	517.8	120.4	-76.8%	Yes
Workload Arrearage ("Batches")[e]	45.78	2.72	-94.1%	Yes

Note. The dimensions of production effectiveness listed were selected on basis of relevance to test-site management's goals of increasing the level, efficiency, and quality of data-transcribing workflow.

[a]Base Period extended from 5 July to 2 October 1976. Period had 2 holidays and 63 work days.

[b]Trial Period extended from 17 January to 16 April 1977. Period had 1 holiday and 64 work days.

[c]Computation used adjusted output (in parenthesis) for Base period, which had one less workday than Trial period. Adjustment added daily rate to raw total.

[d]Man-hours used is, by itself, not an accurate indicator of production effectiveness because the used man-hours can vary considerably for the comparative periods due to number of holidays, amount of annual or sick leave taken, etc. When used in conjunction with other production indicators, however, the number of man-hours used is very important--such as in, for example, determining the efficiency of production in terms of keystroke output per man-hour input.

[e]Measured in terms of the number of swing-shift batches of backlog, on basis of daily average. The comparative ranges of the two periods are as follows: Base Period--High, 76; Low, 23. Trial Period--High, 24; Low, 0. The comparative number of zero-backlog days were 0 and 38 for the Base Period and Trial Period, respectively. The swing shift was the only work shift for which workload backlog records of the base period were available.

Table 3

Test-Site Savings: Actual and Projected

Out Year (end of)	Lump-Sum Savings Compounding[a]			Periodic Savings Compounding[b]			Total Savings Value
	Lump-Sum Savings	Compounding Factor[c]	Compounded Value	Projected Monthly Savings	Compounding Factor[d]	Compounded Value	Lump-Sum and Periodic Savings
1	1,360	1.11	1,510	3,427	12.595	43,163	44,673
3	1,360	1.37	1,863	3,427	42.138	144,407	146,270
5	1,360	1.69	2,298	3,427	78.547	269,181	271,479

Note. Projections for the periodic savings are based on coinciding compounding and series-payment intervals, as described in reference entry for Fabrycky and Thuesen (1974). Interest rate used is 10 percent, as prescribed in DODINST 7041.3.

[a]Actual cost savings, after absorbing set-up costs, recorded during PCRS trial period.

[b]Projected cost savings after PCRS trial period ended, and exclusive of set-up costs--which were absorbed previously.

[c]Assume the actual trial-period net savings were "deposited" one month after that period ended, on lump-sum basis.

[d]Assume first payment of projected production-cost savings had accrued exactly one month after trial period ended, and would recur monthly.

Table 4

Projections of Production-Cost Savings: By Level of Aggregation and Level of Generalizability

Aggregation Level[a]	Projection Factor[b]	Out Year Savings[c]		
		1 YR	3 YR(s)	5 YR(s)
Generalizability Level of 0.50				
Testsite Shipyard	2,622	33K	110K	206K
All Shipyards	20,152	254K	849K	1.6M
NAVMAT Community	73,084	920K	3.1M	5.7M
Generalizability Level of 0.75				
Testsite Shipyard	3,933	50K	166K	309K
All Shipyards	30,228	381K	1.3M	2.4M
NAVMAT Community	109,626	1.4M	4.6M	8.6M
Generalizability Level of 1.00				
Testsite Shipyard	5,244	66K	221K	412K
All Shipyards	40,303	508K	1.7M	3.2M
NAVMAT Community	146,168	1.8M	6.2M	11.5M
Generalizability Level of 1.25				
Testsite Shipyard	6,554	83K	276K	515K
All Shipyards	50,379	635K	2.1M	4.0M
NAVMAT Community	182,710	2.3M	7.7M	14.4M
Generalizability Level of 1.50				
Testsite Shipyard	7,865	99K	331K	618K
All Shipyards	60,455	761K	2.5M	4.7M
NAVMAT Community	219,253	2.8M	9.2M	17.2M

Note. Interest rate used is 10 percent, as prescribed in DODINST 7041.3.

[a]The numbers of data transcribers located at the specified levels of aggregation were as follows: Test-site Shipyard, 26; All Shipyards, 200; and NAVMAT Community, 725.

[b]The Projection Factor is determined by the following product: (Trial Period Production-Cost Savings at Monthly Rate) x (Aggregation Factor) x (Generalizability Factor). As detailed previously in text, the test-site savings in production costs during one full quarter (13 weeks) of operation were $10,281.47, which implies a monthly savings rate of $3,427 after rounding. The aggregation factor is $\frac{N}{17}$, where N represents number of data transcribers at specified level of aggregation, as described in Footnote a; the 17 in denominator represents number of data transcribers used during trial period. Derivation and interpretation of the generalizability factor were described previously in text.

[c]Compounding factors used in the projections are based on coinciding compounding and savings periods (monthly), as described in reference entry for Fabrycky and Thuesen, 1974. Compounding factors corresponding to the 1, 3, and 5-year projections are 12.595, 42.138, and 78.547, respectively. Derivation of the compounding factors is outlined in Appendix E. Projected savings are rounded to nearest unit for the K-valued entries and to nearest tenth of unit for M-valued entries, where K = 1000 and M = 1,000,000.

REFERENCES

Belcher, D. W. Compensation administration. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1974.

Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), The Handbook of Industrial and Organizational Psychology, Chapter 7. Chicago, Illinois: Rand McNally, 1976.

Fabrycky, W. J., & Thuesen, G. J. Economic decision analysis. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1974.

Fein, M. Wage incentive plans. In H. B. Maynard (Ed.), Industrial Engineering Handbook, Ch. 2, Section 6, (3rd ed). New York: McGraw-Hill, 1971.

Flamholtz, E. Human resource accounting. Encino, Ca.: Dickenson Publishing Company, Inc., 1974.

Kazanowski, A. D. A standardized approach to cost-effectiveness evaluation. In J. M. English (Ed.), Cost effectiveness: Economic evaluation of engineered systems. New York: Wiley, Inc., 1968.

Lawler, E. E., III. Pay and organizational effectiveness: A psychological view. New York: McGraw-Hill, 1974.

Likert, R. The human organization: Its management and value. New York: McGraw-Hill, 1967.

Perlman, R. Labor theory. New York: Wiley, Inc., 1969.

REFERENCE NOTES

1. Middendorf, J. W., II. Economic analysis and program evaluation for Navy resource management. SECNAV INSTRUCTION 7000.14B, 18 June 1975.

2. Recognition. Federal Personnel Manual, Inst. 229, Subchapter 3, 21 May 1976.

3. General schedule (GS) for salaried positions. Effective 10 October 1976. Issued December 1976 by Civilian Personnel Office, Classification & Wage Staff, Naval Electronics Laboratory Center (now Navy Ocean Systems Center), San Diego, CA 92152.

APPENDIX A

DERIVATION OF RECHARGE RATE

PRECEDING PAGE BLANK-NOT FILMED

APPENDIX A

DERIVATION OF RECHARGE RATE

The Recharge Rate (RR) is an hourly composite cost per data transcriber that includes (a) basic salary, (b) acceleration, (c) supervision, (d) machines, and (e) overhead. Bases for the current level of cost of each of these components are specified below.

I. Basic Salary of Data Transcriber at Test Site: $4.04

This hourly rate corresponds to the GS-3 Pay Grade, median step, as listed in official General Schedule for Salaried Positions, Effective 10-10-76. The schedule was issued as specified in Note 3.

II. Acceleration: $1.31

This component represents the government share of pension, leave, and other benefits of the data transcriber. The rate is specified by the comptroller's office at the test site. Currently, the rate is 32.5% of basic hourly salary described in I, above.

III. Supervision: $1.59

Each shift has the following supervisory personnel:

Title	Pay Grade	Basic Salary (median step)
Lead Data Transcriber	GS-4	$4.53
Shift Supervisor	GS-5	$5.07
	Total	$9.60

Since there is an average of 6 data transcribers per shift, $9.60/6 = $1.60 per data transcriber. This value differs by a penny from cost level specified by test-site comptroller. No basis for the disparity is known, but it is of minimal significance. Reference for basic salaries of the respective types of supervisors is cited in I, above.

IV. Machine Costs: $0.38

The cost for "CMC" (described in Appendix B) data-transcribing machines is $1180 per month, as specified by test-site comptroller. The number of data transcribing hours in a 30-day month having 3 shifts per work day is determined as follows: 30(5/7) = 21.4 work days, which implies there are approximately 64 shifts per month. Since each shift has 8 hours, there are 512 hours eligible for data-transcribing activities each month. Accordingly, $1180/512 = $2.30 per hour per shift. Finally, since each shift has an average of 6 data transcribers, the hourly machine cost per data transcriber is $2.30/6 = $0.38.

V. Overhead: $3.50

This represents the general-and-administrative-overhead cost charged by the Management Information System Department of all test-site activities that

"purchase" data-transcribing services from the department. The level of overhead cost is determined by the test-site comptroller, and periodically updated. The current rate is $3.50 per hour per data transcriber.

VI. Overall Summary of Recharge Rate Components

Basic salary	$4.04
Acceleration	$1.31
Supervision	$1.59
Machines	$0.38
Overhead	$3.50
Total	$10.82 per hour per data transcriber

APPENDIX B

PCRS BONUS CALCULATION

PRECEDING PAGE BLANK-NOT FILMED

APPENDIX B

PCRS BONUS CALCULATION

I. Terminology

Symbol	Title	Definition	Numerical Expression
* MT	Machine Time	Time spent operating the data-transcribing machines	Time on Computer Machinery Machine (CMC) + time on International Business Machine (IBM).
RR	Recharge Rate	Multiple-component hourly cost of data-transcribing operation, on per data transcriber basis.	Overall cost: $10.82. Components, and the derivation of their respective costs, are described in Appendix A.
SPI	Superior Productivity Increment	Productivity exceeding the "normal" level designated as standard.	PF-1.0. PF represents the Productivity Factor, described below.
* PF	Productivity Factor	Composite factor representing the product of (a) total efficiency of data transcriber across designated data-transcribing tasks, and (b) proportion of time utilized on CMC-machine to time assigned on same.	(a): Total Efficiency = $\frac{\sum \text{Recorded KS}_i}{\sum HR_i} \Big/ \frac{\sum (\text{Standard KS}_i \times HR_i)}{\sum HR_i}$ where HR_i represents hours of operating time on CMC-T, defined below, for task i; KS is # of keystrokes; and (Standard KS_i x HR_i) represents the weighted average of keystroke standards for all tasks i performed by a given data transcriber.
			(b): Utilization Rate = $\left(\frac{\sum HR_i}{\text{CMC-T}} \Big/ 0.75\right) < 1.25$.
CMC-T	CMC-Machine Time	Time assigned to CMC machines in terms of cumulative hours.	(# of work days in accounting period) x 8 + (Overtime hours on CMC-machine) - (Leave Time + Admin Time + Time assigned to IBM).
DTSR	Data Transcriber Sharing Rate	Portion of productivity induced cost savings received by data transcriber .	0.11

* It should be noted that the Productivity Factor (PF) refers to only the CMC machines, while Machine Time (MT) also includes the IBM machines. This reflects the fact that site management wants all data transcribing to be done on the newer CMC-machines whenever possible because they are technologically superior to the older IBM-machines. Thus, only the CMC-machines are used in deriving the Productivity Factor, which is the foundation on which the incentive bonuses are based. On the day shift, however, the workload demands that some work still be done on the IBM machines; in order to permit data transcribers forced to use these machines to be eligible for incentive bonuses, these machines are also included in Machine Time. These interrelationships are illustrated in the basic formula for calculating PCRS bonuses, described below

II. PCRS Bonus Formula

Using the given terminology, the basic formula for calculating the dollar bonus for a specified accounting period is as follows:

Bonus = MT x RR x SPI x DTSR

Functionally, the formula can be meaningfully interpreted in the following sequence:

1. MT represents the cumulative time, in hours, spent on direct production activities for given data transcriber.

2. (MT)RR represents the cumulative cost of the direct production activities per data transcriber.

3. (MT x RR)SPI represents the productivity-induced cost savings for the PCRS site when a given data transcriber has effected superior productivity, i.e., SPI > 0.

4. (MT x RR x SPI)DTSR represents the portion of the PCRS cost savings received by the keypuncher who generated them.

III. PCRS Bonus Computation Example

A realistic example of a monthly bonus might include the following values for parameters:

1. Productivity Factor

   a. Efficiency component: 1.6
   b. Utilization rate: 1.05

2. Machine Time: 150 hours

3. Recharge Rate: Currently $10.82 per hour per data transcriber

4. Data Transcriber Sharing Rate: Currently 0.11

Inserting these values into the basic formula gives the following results

Bonus = MT x RR x SPI x DTSR

= 150 x $10.82 x [(1.6)(1.05) - 1.0] x 0.11

= $121.40

IV. Bonus and Salary Relationships

One way to evaluate the potential utility of the PCRS bonus to the data transcriber is to contrast it with the data transcriber's basic salary. Appendix A indicates that the basic salary of at the GS-3 paygrade (median step) is $4.04 per hour. Evaluating the term, 30(5/7)8, indicates there are 171.44 regular work hours in a 30-day month, which means the basic salary is $692.62 before deductions. Numerically rounding both the bonus value in the example of Part III and the salary gives the following percentage of bonus earnings relative to basic salary:

$$\frac{121\ (100)}{693} = 17.5\%$$

APPENDIX C

NONRECURRING SET-UP COSTS OF PCRS

APPENDIX C

NONRECURRING SET-UP COSTS OF PCRS

I. Equipment

Item	Cost	Quantity	Total
Table	80.00	10	800.00
In-Basket	1.85	30	55.55
			$855.55

II. Software Development

Item	Cost	Quantity	Total
GS-11(5) Programmer	*@S&A($12.31)	300 Man-hours	$3693.00

*S&A: Salary plus "acceleration" (32.5% of hourly salary), as described in Appendix A.

III. Personnel Training and Development

Item	Cost	Quantity	Total
Selected Interviews	GS-3(5) *@RR ($10.82)	20 Manhours	$216.40

* RR = Recharge Rate, a multiple-component hourly cost derived by site comptroller for GS-3(5) data transcribers as described in Appendix A.

Item	Cost	Employees	Man-hours each	Total
Questionnaire Admin I	GS-3(5)@ RR	22	(1.25)	297.55
	GS-4(5)@ S&A ($6.00)	3	(1.25)	22.50
	GS-5(5)@ S&A ($6.72)	3	(1.25)	25.20
				$345.25
Questionnaire Admin II	GS-3(5)@ RR	22	(1.00)	238.04
	GS-4(5) @ S&A	3	(1.00)	18.00
	GS-5(5) @ S&A	3	(1.00)	20.16
				$276.20

III. Personnel Training and Development (continued)

Item	Cost	Employees	Man-hours each	Total
PCRS Introduction	GS-3(5) @ RR	22	(1.00)	238.04
	GS-4(5) @ S&A	3	(1.00)	18.00
	GS-5(5) @ S&A	3	(1.00)	20.16
				$276.20

Supervisory Development of Head of Digital Computer Operations Branch

		Weekly Man-hours	Number of weeks	
PCRS Introduction	GS-11(5) @ S&A ($12.31)	3.5	(1)	43.08
PCRS Training (NPRDC)				
Time	@ S&A	16.0	(1)	196.96
Travel	@ Reimbursed Cost (Auto and Food)			30.00
PCRS Maintenance (weeks 6-13)	@ S&A	10	(8)	984.80
				$1254.84

PCRS - Monitor Development		Weekly Man-hours	Number of weeks	
PCRS - Introduction	GS-11(5) @ S&A	3.5	(1)	43.08
PCRS - Training (NPRDC)				
Time	@ S&A	16.0	(1)	196.96
Travel	Reimbursed Cost (Auto and Food)			30.00
PCRS Maintenance (weeks 1-5)	@ S&A	15	(5)	923.25
				$1193.29

IV. Possible Unrecorded Costs

Add 10% of sum of I, II, III to adjust for possible unrecorded costs.	$811.07

V. Summary of Nonrecurring PCRS Set-up Costs From Test-site Perspective

A. Recorded Costs

Equipment	855.55
Software Development	3693.00
Personnel Training & Development	3562.18
Total Recorded Costs	$8110.73

B. Possible Unrecorded Costs (Estimate 10% of recorded costs) 811.07

Total Nonrecurring Set-up Costs $8912.80

The set-up costs are deliberately given a liberal computation. That is, in addition to actual expenditures (e.g., equipment and travel) there is also inclusion of implied costs due to decreased productivity while data transcribers and direct supervisors were being trained for PCRS implementation during work hours; also included was reimbursement for staff use (e.g., computer programmer who did the software development and, separately, development of the on-site PCRS monitor) and for use of higher level supervision (e.g., Head, Digital Computer Operations Branch). Finally, a substantial adjustment for possible non-recorded costs is also included.

Even given the liberal computation of nonrecurring set-up costs described above, such costs were more than fully recovered during the trial period from the production-cost savings associated with PCRS implementation, as documented in text.

APPENDIX D

COMPARATIVE PRODUCTION EFFECTIVENESS OF WORK SHIFTS

APPENDIX D

COMPARATIVE PRODUCTION EFFECTIVENESS OF WORK SHIFTS

Tables 5 and 6 show the comparative production effectiveness within work shifts across periods. The important summary findings in Table 6 are the following:

1. All shifts substantially decreased in overtime man-hours. As a result, keeping the full complement of data transcribers busy after the workload backlog was essentially eliminated became an occasional problem on all shifts.

2. Grave shift decreased slightly in production efficiency while other shifts increased. Since Grave Shift is last shift for each work day, the decrease might be due primarily to insufficient workload; lack of complete understanding of bonus computation by some data transcribers on this shift may also have influenced the decrease.

3. Swing shift production efficiency increased dramatically: 75.3%. Since tasks, data transcribers, and machines were essentially identical for this work shift across the comparative periods, the increase was presumably attributable primarily to the effects of two factors: (a) the PCRS, and (b) the exchange of day shift and swing shift supervisors that occurred between the periods. Given the nature of the field test, it is difficult to determine the relative impact of these factors on the productivity increase of this work shift.

TABLE 5

Production Effectiveness Across Work Shifts for Base Period and Trial Period

SHIFT[a]	BASE PERIOD[b]				TRIAL PERIOD[c]			
	Key-Strokes[d]	Total Man-hours	Overtime Man-hours	Keystrokes Per Manhour[e]	Key-Strokes	Total Man-hours	Overtime Man-hours	Keystrokes Per Manhour
DAY (N=9)	(14,287,059) 14,063,824	3940.3	222.3	3569.2	14,994.043	3763.6	68.1	3984.0
SWING (N=4)	(6,613,361) 6,510,027	1696.8	36.8	3836.6	11,888,619	1768.0	6.9	6724.3
GRAVE (N=4)	(15,218,433) 14,980,645	2114.7	258.7	7084.1	10,234,551	1562.7	45.4	6549.3
	(36,118,853) 35,554,496	7751.8	517.8	4587	37,117,213	7094.3	120.4	5232

Note. This table is intended to be used in conjunction with Table 6, also in this appendix. That table reflects changes across the comparative periods on a relative basis.

a. All data transcribers were on the same respective shifts for both periods. Number of data transcribers operating under PCRS conditions on each shift is represented by N.

b. Base period extended from 5 July to 2 October 1976. Period had 2 holidays and 63 work days.

c. Trial period extended from 17 January to 16 April 1977. Period had 1 holiday and 64 work days.

d. Computation used adjusted output (in parentheses) for Base Period, which had 1 less work day as explained in Footnotes b and c. Adjustment added daily rate to raw total.

e. Keystrokes per man-hour were computed on basis of raw total of keystrokes - i.e., before adjustment described in Footnote d. The overall average keystroke per man-hour across shifts is, of course, a weighted average with unequal weights.

TABLE 6

Relative Production Effectiveness Across Work Shifts for Base Period and Trial Period

Shift	Inter-period Change: Percent and Direction				Direction of Change Desirable ?			
	Key-strokes	Total Man-hours	Overtime Man-hours	Keystrokes per Man-hours	Key-strokes[a]	Total Man-hours[a]	Overtime Man-hours[b]	Keystrokes per Man-hours
DAY (N=9)	+4.9	-4.5	-69.4	+11.6	yes	yes	yes	yes
SWING (N=4)	+79.8	+4.2	-81.2	+75.3	yes	no	yes	yes
GRAVE (N=4)	-32.7	-26.1	-82.4	-7.5	no	yes	yes	no

Note. This table is intended to be used in conjunction with Table 5, also in this appendix. That table shows the comparative data on which the inter-period changes reflected in this table are based. Inter-period changes across all shifts combined are shown in Table 2, in text.

a. As separate types of indicators, keystrokes and man-hours are only meaningful in terms of "Desirability" in the general sense that more production (keystrokes) is preferable to less and, analogously, lower cost (man-hours) is preferable to higher. In the present case, however, keystrokes and man-hours by themselves are not reliable indicators of production effectiveness because they can vary considerably due to number of holidays, amount of leave taken, etc., within the respective periods. It is only when keystrokes and man-hours are used in conjunction with each other as a composite indicator of overall efficiency, i.e., keystrokes per man-hour, that their interactive importance is primarily demonstrated.

b. Overtime man-hours is an important indicator of production effectiveness in the present case because (1) it reflects the substantial decrease in workload backlog during the trial period, and (2) it represents an excess-cost production penalty due to overtime pay being higher than regular salary.

APPENDIX E

DERIVATION OF FUTURE COMPOUNDED-VALUE FACTORS

APPENDIX E

DERIVATION OF FUTURE COMPOUNDED-VALUE FACTORS

I. Effective Interest Rate Determination

DODINST 7041.3 specifies that the interest rate to be used in economic analysis of DOD programs and end-products is 10%, compounded annually. As explained in text, however, the PCRS cost savings are accrued and compounded on a monthly basis. Accordingly, the nominal annual interest must be converted into its effective annual rate. This is done by the following formula (Fabrycky & Thuesen, 1974; SECNAV INST, Note 1):

$$i = \left[\left(1 + \frac{r}{c}\right)^{c} - 1\right]$$, where i = effective annual interest rate, based on monthly compounding
r = nominal annual rate, based on annual compounding
c = number of annual interest periods

Inserting parameters from the present case, the value of i is determined as follows:

$$i = \left[\left(1 + \frac{0.10}{12}\right)^{12} - 1\right] = 10.47.$$

Dividing effective annual interest rate by number of annual interest periods gives the effective interest rate per specified period. Thus, for the present case, 10.47/12 = 0.87% per month.

II. Determining Future Compounded-Value of Cost Savings

A. Lump-sum Compounding

The basic formula for determining the future value of a single payment that is compounded over a specified number of periods is the following:

$$F = P\,(1 + i)^{n}$$ where F = future compounded value at end of last period
i = effective interest rate per specified period
p = present value of single payment
n = specified number of periods

Sometimes the value of F is already known, but its present value needs to be determined. Then, with the same terminology, the following formula is used.

$$P = F\left[\frac{1}{(i+1)^{n}}\right]$$

B. Series-payments Compounding

The basic formula for determining the future value of a series of regular, equal payments that are compounded over a specified number of interest periods is the following:

$F = A\left[\frac{(1+i)^n - 1}{i}\right]$, using terminology analogous to that described above for lump-sum compounding, and where A represents the amount of each payment in the series.

When the present value of the series payments needs to be determined, the following formula is used:

$P = A\left[\frac{(1+i)^n - 1}{i\ (1+i)^n}\right]$, using terminology analogous to that described above for series-payments compounding.

The future-compounded-value and net-present-value formulae for both lump-sum and series payments are cumbersome to evaluate numerically when the number of periods gets large. Accordingly, references are available (Fabrycky & Thuesen, 1974) which provide equivalent "factors" for frequently-used combinations of interest rates and interest periods. Less-frequent combinations can be derived by interpolating the factors already known. This will be explained in III, below.

III. Derivation of Compounded-Value Factors

The series-payment, compounded-value factors for the 3-year projections shown in Tables 3 and 4 in the text can be derived via interpolation by using the following information from appendices in the reference entry for Fabrycky and Thuesen, 1974:

Effective Monthly Interest Rate	Number of Series Payments	Compounded-Value Factors
1.00	35	41.660
0.75	35	39.854
1.00	40	48.886
0.75	40	46.446

As described in I, above, the effective annual interest rate corresponding to the nominal annual rate of 10% is 10.47% when compounded monthly. Dividing the effective annual interest rate by 12 gives the effective monthly rate: 0.87%. The compounded-value factors associated with this effective monthly rate for interest periods of 35 and 40 an be derived as follows:

(a) For 35 interest periods

$$(41.660 - 39.854)\left(\frac{.87-.75}{1.0-.75}\right) + 39.854 = (1.806)(0.5) + 39.854 = 40.757$$

(b) For 40 interest periods

$$(48.886 - 46.446)\left(\frac{.87-.75}{1.0-.75}\right) + 46.446 = (2.44)(0.5) + 46.446 = 47.666.$$

Then, determine the difference between compounded-value factors for interest periods corresponding to 35 and 40 at the effective monthly interest of 0.87%. Pro-rate the difference in relation to the specific number of interest periods wanted (in the present case, 36), and add to the factor corresponding to lower number of interest periods:

$$(47.666 - 40.757)\left(\frac{36-35}{40-35}\right) + 40.757 = (6.909)(0.2) + 40.757 = 42.138,$$

when rounded in conservative direction.

This compounded-value factor can then be multiplied by the amount of series payment to result in a value equal to that resulting from use of the basic compounded-value formula for series payments:

$$(\text{Series Payment})(\text{Compounded-Value Factor}) = (\text{Series Payment})\left[\frac{(1+i)^n-1}{i}\right] = F,$$

where F represents the future compounded-value of the payment series at end of last period; i = effective interest rate per specified period; n = number of periods. Thus, for the 3-year projections of production-cost savings from 17 data transcribers shown in Table 3, the value of F is determined as follows:

$$(\$3,427)(42.138) = \$3,427\left[\frac{(1+0.87)^{36}-1}{0.87}\right] = \$144,407$$

The process for deriving the compounded-value factors for lump-sum savings, or net-present-value factors of either lump-sum or series-payments savings, is analogous to example given.

ABOUT THE AUTHORS

Dr. Gene E. Bretton is employed by the Management of People and Organizations Program of the Navy Personnel Research and Development Center. He received his BA in Industrial Psychology from the University of Minnesota. His Ph.D. in Organizational Behavior/Industrial Relations was received from the Graduate School of Business Administration, University of California, Berkeley. Among his interrelated research interests are Wage & Salary Administration, Human Resource Accounting, Program Evaluation, and Management Information Systems.

Dr. Steven L. Dockstader is currently assigned to the Management of People and Organizations program at Navy Personnel Research and Development Center. He received his BA and MA degrees at San Jose State University in Experimental Psychology and his Ph.D. from the University of Denver in Human Learning and Motivation. Current research interests are the effects of performance measurement feedback on productivity.

Dr. Delbert M. Nebeker received his BS in Psychology and Sociology from Brigham Young University and his Ph.D. in Social Psychology from the University of Washington with emphasis in Organizational Behavior. Currently he is employed by the Navy Personnel Research and Development Center, San Diego, California in the Management of People and Organizations program. He is a member of the American Psychological Association, the Society for Organizational Behavior, and serves as a member of the editorial board of Organizational Behavior and Human Performance journal. His research interests lie in modeling individual and group productivity and using these models to develop techniques which enhance organizational performance.

Dr. E. Chandler Shumate works in the Management of People and Organizations program at the Navy Personnel Research and Development Center, San Diego, California. He received his BS in General Psychology at Brigham Young University. He received his Ph.D. in Experimental Psychology at the University of New Mexico with emphasis in the area of Human Learning and Cognition. His current research interests are in work motivation and productivity from a learning theory perspective.

EFFECTS OF THE OPERATIONAL ENVIRONMENT ON PERFORMANCE MEASUREMENT

CAPT James J. Clarkin, USN
Navy Personnel Research and Development Center
San Diego, California

ABSTRACT

Assessment of performance in an operational environment can be subject to many difficulties: (1) the mission takes precedence over measurement, (2) nonprogrammed events are very likely to occur, (3) multiple system and subsystem interactions may obscure interpretation of the measures, (4) sample sizes may be small, and (5) opportunities to use sophisticated measurement instrumentation may be limited. Measurement of performance should not be made in the operational environment when it can be performed in the laboratory or in simulation. Nevertheless, there are many circumstances where operational measurement is most desirable despite limitations on the measurement process.

INTRODUCTION

Assessment of human performance in an operational environment ("on the job") presents unique problems to the measurement specialist. In part, this is because certain characteristics of the operational environment make accurate and reliable measurement difficult, and also because there are almost an unlimited number of different operational environments, each of which has its individual characteristics. Among such environments one finds, for example, those on board different types of ships or submarines, in different types of aircraft, in ground control centers, and in many different types of maintenance organizations, both aboard ship and ashore.

In may, if not most, military organizations, personnel are not usually evaluated by precise performance measurement, but by use of subjective techniques such as check lists or ratings, although there are exceptions where hands-on performance actions are evaluated. Harris and Mackie (1962) report that a majority of performance evaluations are made by subjective judgment. In some cases, performance measurement is directed more toward total system or mission effectiveness, and individual performance (whether correctly or not) receives less attention than that of the team or the system. System performance is likely to be measured in terms of end items produced, malfunctions corrected, or targets destroyed; team or crew performance is often evaluated either in terms of system performance criteria or in judged efficiency of the team; and individual performance is usually graded by a supervisor using a checklist, the criterion here being general performance effectiveness rather than specific observed actions. Researchers requiring more precise measurement and observation of interaction effects would in such cases need to set up additional measurement procedures, if indeed this were feasible. Rabideau (1964) suggests that measurements involving many data points and rigid control over variables will be difficult to accomplish in a field setting.

PRECEDING PAGE BLANK-NOT FILMED

Christensen and Mills (1967) have pointed out that impressive differences exist between actual operational conditions and those simulated in the laboratory or training environment. Since many performance-related factors affecting measurement in the operational environment may not be present in the laboratory environment, it is usually assumed (Chiles, 1967; Christensen, 1975[a]; Christensen & Mills, 1967) that a more accurate assessment of performance can be made on the job. There are, however, both advantages and disadvantages to measurement in an operational environment.

OPERATIONAL CHARACTERISTICS AFFECTING PERFORMANCE MEASUREMENT

Some of the characteristics that differentiate the operational environment from the nonoperational (e.g., laboratory, training school, or simulator) situation are as follows:

1. The mission takes precedence over the measurement.

2. Nonprogrammed input contingencies are likely to occur; systems often operate in a degraded mode.

3. System operation includes much subsystem interaction; human action is reflected largely through team operations.

4. Opportunities for task replication are few.

5. Opportunities to use sophisticated instrumentation are limited.

6. Environmental and emotional factors differ from the nonoperational situation.

Each of these characteristics will be described in terms of its implications for human performance measurement. Ways in which problems inherent in these characteristics may be overcome will be discussed.

Mission Takes Precedence Over the Measurement

Perhaps the most important operational characteristic to be considered is the fact that the measurement expert often has little control over many variables important to measurement of operator performance (Rabideau, 1964). The term "operator" as used in this paper refers to any individual who performs system-required tasks, whether the tasks involve the operation or the maintenance of the system.

Such factors as the number of operators performing, the number of researchers or observers required and allowed, restrictions on the use of instrumentation, and changes in work schedules or procedures are controlled by the Commanding Officer (C.O.) or a delegated supervisor of the unit whose performance is being measured, not by the measurement specialist. Thus, details of the organization's mission or assignment to some extent shape the operation of the measurement program, and changes in the mission are likely to affect measurement procedures. The measurement specialist has little control over the specific task, the total system operation, or any environmental or mission variables. He often must measure unobtrusively, exercising care to avoid interfering with on-going work. As Rabideau (1964) puts it, the investigator "is reduced to collecting data on various system

inputs, outputs and intervening processes. In short, he becomes an observer."

Lack of controls present problems for the researcher, some of which are:

1. Work schedules may be changed without notifying the researcher, thus upsetting his schedule and possibly causing him to lose measurement opportunities.

2. In some organizations, there is little time for operators to participate in measurement projects during regular duty hours. This means that operators may be co-opted for testing during their free time. This will probably cause them to view the measurement process less favorably. This situation arises primarily when special tests (e.g. paper and pencil) are required for the operator and would not arise if observations were made of routine work.

3. Operators may not have been trained to perform their tasks in the most efficient manner or in modes desired by the experimenter. There is little likelihood of providing operators with supplemental training that will make measurements more precise.

4. Measurement by observation alone may result in less precision than desired and also requires additional measurement personnel. The opportunity to instrument performance is often lacking.

5. Mission-related interruptions may occur during measurement data collection and thus invalidate portions of the data.

6. Inability to control or manipulate the system variables causes problems not only during measurement data gathering, but also in interpretation of results.

The measurement specialist must expect such problems and take steps to overcome them where possible. Since control of so many relevant factors lies with the C.O., it is obvious that maximum support by the C.O. is highly desirable and can lessen these measurement problems. Prior to any test, efforts should be made to convince management of the importance of the measurement project and the need for its cooperation. Once the management has decided that the project is worthwhile, cooperation of the operating crews will probably be good, since personnel usually take their cues from their superiors. Cooperative operational personnel can be extremely helpful in ironing out scheduling problems, providing the researcher with equipment or operating information, and so forth.

Advance planning can do much to avoid or overcome problems. Anticipation of constraints, such as restrictions to measurement by observation only, or with only a limited number of operators available, can lead the measurement specialist to develop measurement approaches to optimize results. In an operational setting, the researcher must be prepared to modify his measurement plans and schedules to coordinate with the activities and conditions of the organization. In making such adjustments he must, of course, take care not to invalidate his results.

Nonprogrammed Input Contingencies

Researchers must be alert to the effect of nonprogrammed events, such as system malfunction or change in mission goals, upon the measurement program, since

procedures may be altered or crew members shifted in duties without notice to the researcher. Rabideau (1964) mentions that the result of nonprogrammed events may be new man-machine interfaces, procedures, and human performance requirements. The effect upon measurements of task interruptions or changes in system operation are difficult to evaluate but must be considered when determining the reliability or validity of the measurements.

If system operation is degraded, some part of the tasks may change from automatic operation to manual, some personnel interactions may change, procedures may be altered because of missing inputs, and so forth.

Such nonprogrammed occurrences are not, however, completely disadvantageous. Presumably they arise from genuine mission requirements and, therefore, reflect the reality of system functioning. The researcher should be prepared to collect data on these occurrences, their relation to changed system requirements, and their effect on personnel performance. Moreover, they provide an opportunity of identifying additional measurement needs not apparent in the laboratory situation, such as the effects of stress and/or degraded system operation.

The disadvantages resulting from such occurrences are that measurements in progress may be lost, measurement time may be lost by rescheduling, and operator morale may be adversely affected. Also, if a degraded condition should continue, with resulting considerable alteration of procedures or man-machine interactions, the entire measurement project may have to be rescheduled.

Probably the only way to deal with problems brought about by non-programmed events is for the researcher to prepare himself for these in advance. It is unlikely that he can change anything in the job situation, but if he understands the system and its interactions and has planned to include alternative measurement methods in case of disruptions, he has a better chance of overcoming the problem.

System Operation Includes Much Subsystem Interaction

When there are many man-machine and man-man interactions within a system, measurement of individual performance becomes more difficult. Keenan (1965) lists the interactions within a system as man-man interaction, interaction of system personnel with the various system equipments, interaction of system personnel and the ambient system conditions (e.g. illumination, sound, vibration, and temperature), and interaction between system personnel and the system activities, procedures, and doctrine. The performance of almost any task is probably affected by one or more of these interactions. Moreover, the operator may interact with equipment and/or personnel in more than one subsystem; therefore, in many cases it will be essential to measure both the individual performance and the team, subsystem, or system output or performance. Doubtless these interactions may have important effects on measures of worker performance, although as Meister (1974) has noted, "It cannot . . . be assumed that every variable influencing individual performance will have a critical effect on the system." Failure to include all significant system interactions in the measurement situation may, of course, reduce the validity of the results.

Since human performance in complex systems is affected by these interactions, measuring the performance of an individual operator in a laboratory, isolated from interactions with other equipment and personnel, probably yields less valid data than measurement in the operational situation. Since the operational

environment is distinguished by such interactions, the researcher's failure to measure them means that he has degraded his measurement to the level of the laboratory.

As an example of multiple interactions, consider a shipboard Combat Information Center (CIC). The CIC operator is a part of so many different man-machine and man-man interactions -- both internal and external -- that any measurement of his performance must consider the effects of all significant interactions (e.g., communication with other operators, reading of displays on equipments other than those he controls, etc.). Indeed, there may be so many interactions that it may be difficult to establish the precise bounds of the tasks to be measured.

The larger the system, the more the number of observers and/or measuring instruments required to cover personnel and interactions within the system. It may sometimes be desirable, for example, to observe all members of a crew and all displayed information simultaneously when system activity occurs.

It may also be difficult to isolate a specific task or factor. In large and complex systems the effect of an operator's performance on system output may be confused because of these multiple interactions. Nevertheless, wherever possible an attempt should be made to isolate this effect, since the significance of personnel performance will be obscure unless it is measured against mission success. Because measuring performance in complex systems is not easy, it becomes increasingly important for the measurement specialist to attempt to understand every phase of the system operation. It may require much time and experience on the part of the research team to obtain the understanding of the ship routines, equipment functions, etc., necessary for efficient performance measurement planning. Without this understanding, however, much time may be lost and results may suffer. In a report on evaluation of the Navy P-3 aircraft Pilot and Tactical Coordinator operator performance, Matheny, Patterson, and Evans (1970) stress the importance of such an understanding.

Opportunities For Task Replication Are Few

Sometimes, because of schedules, workload, or interruptions, it will not be possible for the measurement specialist to get the desired number of repetitions of a certain task, thus lowering the validity of his conclusions. It may not be possible to obtain adequate measurement of performance on some tasks in the operational environment because the tasks occur so infrequently or, as Christensen (1975[b]) suggests, because of safety or cost consideration. Examples include the launch of a nuclear weapon, the tracking of a Soviet submarine (in certain geographic areas), and the use of emergency ejection procedures in an aircraft. Tasks such as these can often be measured more efficiently in a simulator.

The number of operational personnel whose performance can be measured may also be limited. In the most extreme case only one or two operators may be available for measurement. How does one determine if their performance is representative of all operators doing the same job? Or, suppose one has only one measurement each from several operators. Even if the researcher is fairly confident of the validity of his data, the reliability of his results may be questionable because of a small sample size.

There are ways in which the adequacy of such results may be improved. For example, a panel of "experts" could be asked to judge whether the results appear to be a reasonable representation of average or usual operator performance. Or, possibly, the data may be combined with other related data to yield useful results. Of course, such combination and interpretation of data must be done with care.

Nonetheless, although at times it may be possible to accumulate comparatively few measurements of task performance, measurement under operational conditions will still be more valid than the composite of many measurements made in the laboratory. The initial decisions to be made by the researcher are (1) how important are operational factors to the performance to be measured, and (2) will testing in the laboratory or a simulator yield data as satisfactory as those secured from the operational environment?

Opportunities To Use Sophisticated Instrumentation Are Limited

Those accustomed to using laboratory instrumentation to measure such variables as reaction time, error rate, and alertness level may be somewhat frustrated in many operational settings because of reduced opportunities to instrument data collection. Mobile timing and recording equipment are probably permitted in most organizations; however, it is unlikely that the data collector will be allowed to connect such equipment to operating systems because of (1) possible interference with mission operations, (2) the possibility of resultant equipment malfunctions or system errors, (3) operator distraction by the instrumentation, or (4) the system down time involved in installation and checkout of the instrumentation. The restriction on instrumentation means additional emphasis on observational methodology, and this in turn produces problems for the researcher.

Observational methodology is highly subjective and requires special precautions if valid data are to be secured. The observer must be trained to observe the particular behaviors to be measured; his own performance as an observer must be measured and calibrated. This may pose relatively little difficulty when the observed behaviors are relatively objective and discrete (e.g., control panel operation); more severe difficulties may arise when the behaviors are complex and continuous (e.g., tracking). The observational burden increases the number of high level data collection personnel needed; the later are often in short supply. Although it may be possible to make use of operational personnel as observers of their co-worker's performance, in most cases the researcher will have to provide his own observers.

Any answer to these problems depends upon detailed planning of the measurement task. As a minimum a comprehensive analysis will be required of the tasks to be measured, operator interactions and machine events. These must be plotted as a function of time. Training of observer personnel and practice in performing observations are mandatory. Such planning will at least suggest the dimensions of the measurement problem and ways of overcoming it.

Environmental And Emotional Factors

Environmental and emotional factors which impact the performance of operators on the job are often not present in the laboratory or even in a highly realistic simulator. These factors are often quite different from those in the laboratory or training environment. Aboard ship, for example, there may be extremes of noise, vibration,

temperature, and motion that are not usually encountered elsewhere. Performance may change with variations in these environmental conditions.

Psychological factors are less stable in some operational settings. For example, motivation and consequent productivity are likely to vary considerably with differences in the command climate. On a ship or submarine, the operator's entire "world" is different from his world ashore. Work and rest areas are often crowded. Morale will be low at times. Stress and work pressures found in the work situation may not exist in a training or simulator situation. The individual is working for a supervisor or crew leader whose approval is more necessary to him than that of the researcher. Because of mission requirements, the demand for faster or more accurate performance will often be much greater than in the research laboratory, where there is usually no punishment or criticism for poor performance. It has also been found (Chiles, 1971) that the worker tends to work harder at what he considers more important tasks, and thus tasks performed concurrently may show some degradation. Surveillance by supervisors, a familiar environment in which to perform, and operator concern about the effects of his errors affect personnel performance, sometimes positively and sometimes negatively. These factors, often present on the job, are not usually found when measuring in a laboratory setting. Fleishman, Levine, and Glickman (1973) state that "a person's motivation and general behavior are likely to be different in test and work situations, and so are his apparent reactions to stressful situations."

The operational environment is rich in behavioral impact because many factors that may affect job performance are present: physical environment, equipment, crew interactions, interruptions, supervisor pressures, psychological climate, and real or imaginary dangers. Although it may be difficult to obtain meaningful measures of some of these variables, it is desirable for the researcher to judge the degree of each pertinent factor present during a measurement study, because such data may serve to explain otherwise discrepant findings and may be useful in comparing results with those of other measurements.

Even in the operational setting, of course, observer presence and the visibility of the measurement process may affect the operator performance. However, even with this contamination, motivation and stress level should be more like that in a "normal" job performance than would be the case in a laboratory or training situation.

Although laboratory measurement of performance using mockups or single items of actual equipment (e.g., a console) may at times be necessary or desirable, the operational factors described may not be present; thus, the performance measured will not be truly generalizable to the operational work situation. Chiles (1967) says that

> When one attempts to answer questions about human performance as it occurs in operational situations, one becomes painfully aware of the inadequacies of the extrapolations that must be made in attempting to apply research data to the practical problems of the real world.

ADVANTAGES AND DISADVANTAGES OF THE OPERATIONAL ENVIRONMENT

Not all operator performance measurement should be performed in the field. Some tasks may be so simple (e.g., sorting of records) that valid measurements may be made effectively and with less effort or expense in a laboratory setting. With other tasks it may be necessary, or more feasible, to measure the performance in a simulator.

Even with the use of simulators, some operational factors affecting performance may be missing. Some factors, such as extremes of motion, limited work areas, or realistic mixes of noise and signal are extremely difficult to simulate effectively, or are so expensive to simulate that decisions must be made as to how important each increase in realism is. As has been suggested here and by others (Christensen, 1975[b]; Harris & Mackie, 1962; Rabideau, 1964), however, if the task is complex, with many interactions, and the operational environment is judged to greatly affect personnel performance, then measurements should be made in the operational setting.

In measuring performance on the job, there is an opportunity of identifying additional measurement needs and considerations which may be overlooked in the laboratory. System interaction effects also can be observed or measured.

Differences in personnel output are most readily seen and measured directly as individual performance, but, when performance can be observed in operating systems, the significance of personnel outputs can be determined in terms of their effect on overall subsystem or system output. In the end, of course, each measurement problem must be considered individually in light of its measurement requirements, the operational conditions that apply to that problem and the relative advantages and disadvantages of the operational environment. Among factors to be considered are: How complex is the task? What are the personnel and system interactions? What are the organizational and environmental conditions? What are the cost considerations? How much realism is really necessary? If the answers to these questions suggest operational measurement, detailed planning of the operational measurement can begin. Among the elements to be included in that plan are the following:

1. Efforts should be made to convince management (the Commanding Officer or his staff) of the relevance of the project and the probable value of the results to his organization and to the Navy. Concern for noninterference with the mission should be emphasized. Such presentations should be made by the researcher in person.

2. Before beginning the measurement program, efforts should be made to establish rapport with the personnel to be observed or measured in order to reduce any latent resistance or hostility toward "foreigners." The fact that measurement will not adversely affect the personnel being studied should be emphasized. Obviously, results will not be valid if workers are uncooperative.

A knowledgeable naval officer (with as high a rank as possible) on the measurement team will be very helpful in gaining acceptance by operational personnel. Personnel should be promised that they will receive feedback from the measurement.

3. Results and conclusions should be fed back in some simple and meaningful form to the C.O. of the organization as soon as is feasible. This step may be helpful in securing cooperation in the future, especially if some of the results are immediately useful to the Navy.

4. Plans should be made in as much detail as possible to overcome expected operational problems. Rabideau (1964) suggests that alternate plans be developed to handle anticipated contingencies. What will be done, for example, when nonprogrammed contingencies occur, such as mission or weather changes, malfunctioning equipment, or transfer of the worker to another task? How much extra time must be allowed to secure the required number of subject or task replications? How much additional analysis of the equipment, task, and system will be required at the job location to be sure all tasks, crew interactions, etc., are considered?

When the potential effects of operational factors are analyzed in advance and arrangements made to compensate for these when possible, and when all possible steps are taken to secure necessary cooperation from all quarters, a well-designed measurement study performed in the operational environment has an excellent chance of yielding valid and -- what is more important -- useful results.

Lack of control problems

1. work sched may be changed w/o notifying researcher
2. operator may be required to participate testing during other free time
3. operator may not have been trained to perform task
4. may not be possible to use standardized
5. mission related interruptions
6. lack control of extraneous variables - interpretation of data

REFERENCES

Chiles, W. D. Methodology in the assessment of complex performance: introduction. Human Factors, 1967, 9(4), 325-327.

Chiles, W. D. Complex performance: the development of research criteria applicable in the real world. In W. T. Singleton, J. J. Fox and D. Whitfield, Measurement of Man at Work, London: Taylor and Francis, 1971.

Christensen, J. M. Comments on simulation. Proceedings of the 19th annual meeting of the Human Factors Society, October, 1975a.

Christensen, J. M. Human factors in engineered systems--where the (inter)action is. Professional Paper, Aerospace Medical Research Laboratory, WPAFB, Ohio, 1975b.

Christensen, J. M. & Mills, R. G. What does the operator do in complex systems, Human Factors, 1967, 9(4), 329-340.

Fleishman, E. A., Levine, J. M. & Glickman, A. S. A program for research on human performance. Washington, D.C.: American Institutes for Research, June, 1973.

Harris, D., & Mackie, R. Factors influencing the use of practical performance tests in the Navy (TN-703-1). Human Factors Research, Inc., August, 1962.

Keenan, J. J., Parker, T. C. & Lenzycki, H. P. Concepts and practices in the assessment of human performance in Air Force systems (TR-65-168). Aerospace Medical Research Laboratory, WPAFB, Ohio, September, 1965.

Matheny, W. G., Patterson, Jr., G. W. & Evans, G. I. Human factors in field testing (LS-ASR-70-1). Life Sciences, Inc., May, 1970.

Meister, D. Where is the system in the man-machine system? Proceedings of the 18th annual meeting of the Human Factors Society, October, 1974.

Rabideau, G. F. Field measurement of human performance in man-machine systems. Human Factors, 1964, 6, 663-672.

ABOUT THE AUTHOR

Captain James J. Clarkin, USN, was born on 2 January 1931 in New York City. He graduated from Villanova College and attended Columbia Law School prior to being commissioned an Ensign in June 1953. After receiving his commission, Captain Clarkin served on board the USS JUNEAU as Navigator and CIC Officer. Subsequently he has served as Executive Officer of Headquarters Signal Organization at CINC South, Naples, Italy; Operations Officer and Navigator of the USS SARSFIELD; Commanding Officer, USS SHRIKE; Instructor at the U.S. Naval Academy; student at the Armed Forces Staff College; Executive Officer, USS KRAUS; Commanding Officer, USS LESTER; Commanding Officer, guided missile destroyer, USS STRAUSS; and Special Assistant to the Chief of Naval Personnel. During his service career, Captain Clarkin attended George Washington University, where he earned a Masters's Degree in 1964, and Harvard Business School, from which he received a doctorate in 1971. Captain Clarkin's personal decorations include the Bronze Star with Combat "V," the Navy Unit Citation, and the Legion of Merit.

THE HUMAN SIDE OF PERFORMANCE MEASUREMENT

Laurie A. Broedling
Navy Personnel Research and Development Center
San Diego, California

ABSTRACT

This paper deals with the "human" element in performance measurement; that is, those considerations which make the measurement of humans uniquely different from measuring animals or mechanical devices. Described are three categories which pertain when the object being measured is human: legal, ethical, and psychological. Within the first two categories are various constraints on the process of human performance measurement. Psychological considerations include those aspects of the measurement process which may affect people's attitudes and motivation, in turn affecting their performance and productivity. Also described are the considerations which pertain when humans measure the performance of other humans. In such instances, a social situation exists, and what occurs is discussed in terms of current theories of social interaction and interpersonal perception. The primary conclusion is that these "human" elements in performance measurement are worthy of much more consideration than they normally receive and, in addition, that subjects frequently play an active rather than passive role in the process.

In 1960 a landmark book was published called The Human Side of Enterprise, written by Douglas McGregor. This book signalled a major shift in interpretation of the behavior of people at work. In essence, this shift was away from viewing employees as reactive and toward viewing them as desirous of being proactive. Prior theories of work motivation were of two major types: Frederick Taylor's scientific management, which was predicated on the notion of inducing productivity through economic rewards, and the human relations movement, which was predicated on the notion of inducing productivity through social rewards. Both assumed people to be basically passive and resistant toward work; therefore, management must induce, manipulate, or coerce employees into working hard. McGregor instead suggested that human beings are basically proactive and will take an active interest in their work, especially if given a chance to fulfill higher order needs such as building self-esteem. He characterized the difference between people being basically passive vs. active as Theory X vs. Theory Y. This switch in emphasis has revolutionized the study of people in organizations, including human performance measurement. Therefore, the title of McGregor's book has been paraphrased as the title of this paper.

A simplified way of classifying performance measurement situations is into four cells, as shown in Figure 1. This paper is focused on the cell containing the X; that is, situations in which (1) the performance being measured is that of humans as opposed to that of machines or animals, and (2) the measurement is being taken by a human (who may be using some recording device) as opposed to a strictly mechanical recording. This paper is divided into two major sections. One section contains the performance measurement issues which pertain when a human is being measured. The second section contains the issues which specifically pertain when humans are measuring humans.

What is being measured	What is doing the measuring: Human	What is doing the measuring: Non-Human
Human	X	
Non-Human		

Figure 1. Classification of performance measurement situations.

PERFORMANCE MEASUREMENT OF HUMANS

Measuring humans is qualitatively different from measuring animals or inanimate objects because human awareness of being measured can significantly affect both the process and the outcomes. This awareness has effects in three areas: legal, ethical, and psychological. Legal considerations include those laws and rules which regulate how performance can be measured. Such regulations are usually the result of objections raised by people measured in the past. Ethical considerations are those which, while not codified into law, are commonly felt worth observing in accordance with moral principles. Psychological considerations pertain to the effects which performance measurement has on the mental state of the people being measured; that is, on attitudes and motivation. Changes in attitudes and motivation, in turn, can have strong effects on performance and productivity. Ideally, one would like to use the process of performance measurement as a positive force in productivity enhancement. At the very least, one would like to ensure that the process of performance measurement has no negative effects on productivity.

Legal Considerations in Human Performance Measurement

There are a number of existing regulations which constrain the process of human performance measurement. Among the most important federal laws are the Privacy Act, the Freedom of Information Act, the Civil Rights Act, and the Department of Health, Education, and Welfare (HEW) guidelines on the protection of human subjects.

The general intent of the Privacy Act is to restrict the types of information which can be collected on individuals and to require a description of the purposes of gathered information. With regard to research, the Privacy Act attempts to provide the public with the opportunity for informed consent. The Act makes it unlawful for federal agencies or contractors to maintain records on individuals without the public being aware of the existence of such record systems. Therefore, the existence of a system of personal records must be published in the Federal Register. The applicability of the Act to the Navy is described in SECNAV Instruction 5211.5. The Act also makes it possible for individuals to see most of the records kept on them, including performance records. The Privacy Act includes penalties against violation, including a maximum $5,000 fine. Violations include (1) willfully disclosing individually identified information

to anyone not entitled to see it, (2) not publishing in the Federal Register the existence of a system of personal records, and (3) obtaining records under false pretenses. The fine can be levied against anyone regardless of their institutional affiliation; that is, collecting information as an employee of an organization is no protection against individual liability. Since the provisions and applicability of the Act are somewhat vague, people generally have been very conservative in interpreting it. For example, some civil service supervisors have suspended keeping critical incident lists of employee performance for appraisal purposes in case such lists might be in violation of the Act. However, the Civil Service Commission has issued an interpretation which says that supervisors' notes are not necessarily agency records for purposes of the Privacy Act ("Supervisor's Notes OK," 1976).

The general intent of the Freedom of Information Act is to increase the availability of information on individuals. It requires that government agencies publish in the Federal Register the existence of many types of information which they have and to make such records available to any person requesting them. The burden of proof for withholding information lies with the agency. Ironically, then, the fundamental purposes of the Freedom of Information Act and the Privacy Act are in conflict (Bryant and Hansen, 1976).

Another relevant set of legal considerations pertains to the area of equal employment opportunity (EEO), which first took on importance with the enactment of Title VII of the Civil Rights Act of 1964 (Beach, 1975). Prior to this legislation, there were few restrictions on how employers hired or promoted people. The Equal Employment Opportunity Commission (EEOC), established by Title VII, threw up the first roadblocks by insisting that hiring and promotion decisions should be based solely on measures which were empirically demonstrated to be related to job performance. The landmark Supreme Court decision which backed up the EEOC guidelines in this matter was the case of Griggs vs. Duke Power Company. When this case was decided in 1971, Chief Justice Burger wrote, "What Congress has commanded is that any tests used must measure the person for the job and not the person in the abstract."

Since 1964, the legislation and guidelines pertaining to EEO have proliferated. For example, age has also been specifically excluded as a basis for personnel actions (Age Discrimination Act of 1967). Moreover, many characteristics cannot be used which unduly discriminate against the hiring of minorities, even if there is a demonstrated relationship to job performance (e.g., dishonorable military discharge). With the passage of the Equal Employment Opportunity Act of 1972, the coverage of Title VII was expanded to encompass many more employees (e.g., ones who work for state and local governments) and the Equal Opportunity Commission was given power to institute civil actions in federal courts to ensure compliance with Title VII. With this proliferation of EEO legislation has come a proliferation of confusion on the part of employers about what information they can use to take personnel actions. The courts have not helped clarify matters and, if anything, have added to the confusion, with different courts rendering seemingly contradictory decisions (Sharf, 1977).

One of the results of these EEO developments has been to significantly enhance the importance of performance measurement. Because little reliance can now be placed on demographic and background characteristics in taking personnel actions, attention must be focused exclusively on job experience and work performance. Thus, the need to develop good technology for performance measurement is more pressing than ever before.

Union regulations also can constrain the process of performance measurement, although the specific stipulations vary from contract to contract. In general, unions are opposed to the process of merit promotion and instead advocate promotion based on seniority (Beach, 1975). Therefore, they tend to be wary of individual performance measurement and try to keep it at a minimum. Not only are unions concerned with the procedures for routine performance measurement, but, ordinarily, they must also be consulted if a research study is being conducted entailing performance measurement. Under existing Civil Service regulations, the extent to which federal civil service unions have a voice in whether research studies can be conducted on civil servants (all types of studies, performance measurement included) is undetermined because no regulations specifically address this point. Consequently, in many instances, research projects have been reviewed by local union representatives and, in some instances, have been vetoed by these officials. However, a change in Civil Service regulations has now been proposed in order to clearly stipulate that decisions regarding the conduct of research studies on civil servants is the exclusive right of management. The granting of permission to do performance research on military personnel is a command prerogative. Whether military unionization would change this is a matter of speculation but, given that civil service unions will probably not share this prerogative with management in the future, it is unlikely that a military union would either.

The general intent of the HEW guidelines on the protection of human subjects is to ensure that no abuses or damage occur to those people who serve as research subjects. While the controversy which engendered these guidelines resulted from abuses in medical research such as experiments on fetuses, the guidelines pertain to all research with human subjects, including psychological research. The result has been that most institutions receiving federal research funds have set up committees to review all proposed studies using human subjects in order to prevent abuse of the participants.

Ethical Considerations in Human Performance Measurement

Ethical considerations are those which, while not codified into law, are dictated by common sense and by a general consensus of what is reasonable to do in measuring performance. When these considerations are consistently violated by those who measure performance, then laws usually are passed to formalize public protection. The National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research was formed in response to a widespread perception of incidents of serious ethical abuses on the part of researchers. While the bulk of these abuses have occurred in the medical profession, psychology has by no means been exempt. For example, in psychology's time-honored tradition of dustbowl empiricism, a common approach to developing selection and advancement tests was to measure as many things as possible and then to determine which of them held up under validation and cross-validation. This entailed obtaining some information which seemed to job applicants or employees to be either embarrassing, an invasion of personal privacy, or without face validity. Examples include questions regarding birth control practices, political opinions, and the use of lie detectors.

What is ethical, of course, is a complex matter and varies with the time and the situation (Beals, 1960). The two most frequently raised ethical issues pertaining to studies of humans are the right to informed consent and the right to privacy (Parsons, 1969). With respect to the right to informed consent, the issues are

muddy (e.g., What constitutes being informed?). How can one inform children or the mentally ill? How can the investigator avoid being persuasive and biasing the individual's decision to participate? With respect to the protection of privacy, there is a need for a proper balance between the individual's right to privacy and society's need for information on its members in order to formulate useful social policies. Exactly where this balance is struck depends upon a number of factors (Feldman, 1976). For example, when a society is threatened or in a state of war, the balance will shift toward society's need for information and away from individual rights.

The issue of ethics usually emphasizes the potential risks to the subjects. What is often overlooked is the other side of the coin: the opportunity for the volunteer to make a positive contribution to society. There is an inseparable relationship between risk and progress. For example, new aircraft development would be impossible without people willing to take the risks of flying new, untested types. Therefore, it may be just as unethical to deny people the right to take risks and thus participate in the forward movement of civilization as it is to force them to take unwarranted risks (Edsall, 1969).

Psychological Considerations in Human Performance Measurement

Psychological considerations include the effects which performance measurement has on employee feelings, attitudes, and performance. The fact that the object being measured has feelings and attitudes which affect its behavior is what qualitatively distinguishes the social sciences from the natural sciences. In fact, the social sciences have come under intense attack for overlooking this critical distinction and trying to model themselves after the natural sciences. According to Moore and Anderson (1962), social scientists have frequently tried predicting human behavior as if it were some inanimate process such as the weather:

> What seems frequently to be overlooked is that though the weather is unlikely to change its character <u>because</u> we have made a prediction about it, people, on the other hand, are quite likely to change their behavior <u>because</u> we have predicted that they will behave in such a way . . . More to the point, it seems plausible to suppose that if people are given a reasonably good theory, which enables them to cope conceptually with a broad class of problems, they will <u>use</u> it, and, in that respect, at least, alter their . . . behavior. (pp. 237-238)

In other words, the "laws" of human behavior may actually change as a result of people's awareness of what social scientists have found.

There are three major purposes for measuring performance: (1) to do research, (2) to inform and motivate employees to improve their job performance, and (3) to decide on personnel actions, such as promotions or pay raises. These three can be regarded as on a continuum, with research being the least threatening to the person being measured because it has the least direct impact, and personnel actions being the most threatening. Therefore, depending upon the reason for performance measurement, the psychological considerations differ.

The major psychological considerations pertaining to research fall under what is called reactivity to research. This topic area has been of intense interest to psychologists in the last decade (Rosenthal and Rosnow, 1969). The major conclusion is that experimental results are often contaminated by the subjects' reactions to the experimental situation. The most frequent form of reactivity is that the subject complies with the demand characteristics of the experiment. In other words, the subject ascertains what the experimenter is trying to prove and acts accordingly in order to be compliant or to be helpful. In addition, a subject is more likely to behave in a socially acceptable manner when in an experiment. To avoid the problems accompanying subjects' reactions to being measured, there has been much interest in the use of unobtrusive measures--where subjects are unaware they are being studied (Webb, Campbell, Schwartz, and Sechrest, 1966). Despite the methodological advantage of unobtrusive measures, however, their use raises delicate legal and ethical issues.

The major psychological considerations in informing and motivating employees to do a better job revolve around the giving of feedback. Despite the self-evident importance of feedback, application of feedback principles has been surprisingly neglected in the world of work (Mosel, 1961). On the feedback issue, it is the motivation component that distinguishes measurement of humans from measurement of inanimate objects. The efficacy of a feedback loop to assist a mechanical device in improving its performance must be evaluated only in terms of the feedback's informational component. In the case of humans, however, the feedback loop must be evaluated not only in terms of the usefulness of its information but also in terms of its impact on the person's motivation to perform. Unfortunately, sometimes feedback with high informational value has negative effects on motivation. This situation is most likely to exist when a person is first learning a skill. Early in the learning process people tend to make frequent mistakes, yet this is the period when they are most likely to need the encouragement of knowing that they are doing some things correctly. Therefore, at this point it is more important to feed back information about correct responses and to deemphasize feedback about errors (Mosel, 1961). Moreover, in general, criticism is not an effective motivator (Mosel, 1961; Meyer, Kay, and French, 1965). One of the mistakes most frequently made by supervisors is to discuss the employee's weaknesses and shortcomings; instead, the supervisor should work to build on the employee's strengths (Rogers, 1975). The development of behaviorally based performance appraisal methods are one effective way to give feedback (Kearney, 1976). These are methods which focus (1) on an employee's behaviors and do not address personality traits, and (2) on behaviors which are under the individual's control. The latter is important because many factors which restrict productivity are environmental ones which the employee cannot modify.

The third reason for performance measurement is to take personnel actions. This area is the most threatening to the employee because personnel actions can impact on the employee's self-concept and on his or her image among coworkers. Consequently, the effects on the employee's motivation and performance can be drastic.

Unfortunately, in most organizations, the performance appraisal program entails giving the employee feedback about job performance and taking personnel actions at the same time, usually in an annual appraisal session. This attempt to kill two birds with one stone usually means neither purpose is well-satisfied (Kearney, 1976). Mustafa (1969) has recommended that the Civil Service Commission encourage

agencies to use the required formal appraisal process only for taking personnel actions and to design other, independent means for counseling and developing employees. Research has shown that when both steps are taken simultaneously, the employee listens only for the personnel actions (e.g., pay raises, awards, etc.) and pays little or no attention to the performance counseling aspects (Meyer, Kay, and French, 1965). Therefore, personnel actions should be clearly separated in time from feedback sessions. This separation is even more critical in today's climate of shrinking opportunities for promotion, a situation which is particularly acute in the federal government. Because of the past extensiveness of promotion opportunities, many employees still equate development and performance improvement with eligibility for promotion. Unfortunately, in the present climate of austerity, this is simply no longer the case. One way of dealing with unrealistic employee expectations in these matters is to clearly separate, both in time and in method, actions taken to decide on promotions, lay-offs, pay raises, etc., from actions taken to coach employees on their job performance.

Another reason for separating the feedback component from the official evaluation component in the federal government is related to the poor quality of the latter. For both civil servants and military, the ratings given simply do not discriminate among different qualities of performance. In the case of civil servants, almost everyone is "satisfactory," while in the case of military personnel, almost everyone is "excellent" or "outstanding." Since this administrative rating process is considered farcical by most people involved, it is a mistake to expect employees to pay attention to performance counseling which is given as part and parcel of this process. While some counseling is required as part of the rating process, it is best for supervisors to take some other time to accomplish meaningful performance counseling.

A relatively new concept relevant to performance measurement for personnel action purposes is Human Resource Accounting (HRA). HRA is a method for measuring and quantifying employees' value to the organization. Traditional accounting procedures treat human resources as an expense and record only nonhuman investments. HRA considers some aspects of human resources as assets and supplements conventional financial statements with the changes in net worth of the human resources over the accounting period. HRA makes it possible for management to actually do something about their favorite statement, "Our people are our most important asset." HRA makes intuitive sense as well. An organization's human resources can be quickly dissipated by a hard line manager who exacts productivity by any means. A financial statement which mirrors only direct productivity and nonhuman investments will not accurately reflect what is happening in such an organization: employees' morale and organizational commitment will be dropping, ultimately resulting in (1) the loss of people with valuable experience and expensive training, and (2) a drop in productivity.

In an HRA system, good methods of performance measurement become very important. Some rather detailed procedures have been worked out for HRA (Flamholtz, 1974). However, in spite of the available techniques to do HRA, it has not been used extensively. Only one full-scale use has been made in industry by the R. G. Barry Corporation. There has been some resistance to the cost of carrying out HRA. There has also been resistance to the idea of placing a dollar figure on an individual's worth. Along the same lines, there could be potentially demotivating effects if employees knew what their net worth was to the organization.

Presumably, under the Freedom of Information Act, such information would be accessible to employees. Possibly, HRA will be more accepted if it is carried out on groups but not on individuals.

HRA may find a particularly sympathetic audience in the military due to the military emphasis on evaluating organizational improvement primarily in terms of "hard" indicators, such as productivity or readiness indices. Anything "soft," such as morale, is viewed as suspect because the nature of its impact on readiness or productivity has not been quantified. To the extent that HRA methods convert these "soft" indicators into dollars and cents cost projections of impact, they will probably be seen as highly useful by military management. One need here is to determine exactly what constitutes desirable long-range impact or outcomes for the Navy. Along these lines, a preliminary attempt has been made to develop a personnel status index for the Navy (Borman and Dunnette, 1974). Using a policy-capturing method with Navy officers, three components were important indicators: retention rate, discipline, and readiness.

Another psychological consideration in human performance measurement has been touched on above--whether to take measurements on an individual basis, on a group basis, or on an organizational basis. There are advantages and disadvantages to each. Some of the disadvantages to individual measurement have already been described, such as subject reactivity or threats to the subject's self-concept. Moreover, unions are generally staunchly opposed to management maintaining individual performance records, especially if there is any chance that such records will be used for performance appraisal. Advantages include that it is usually easier from a technical point of view to devise meaningful performance measures for individuals than for groups. Another option is to take individual measurements but then aggregate the measures into group scores and discard the individual data. An advantage of taking direct group performance measures, rather than aggregating individual measures, is that oftentimes, in a Gestaltist way, a group performance is something other than the simple sum of individual performances. For example, in a group consisting of several high achievers but in which only limited individual recognition is available, each person's output might be high, but their attempts to undercut one another to gain the limelight might result in poor group output. When measuring the performance of groups or of organizations, the relevant concept is organizational effectiveness. One of the biggest problems in measuring organizational effectiveness is in obtaining accurate, quantitative records of group output. Rarely do accurate records exist as part of ongoing reporting systems in organizations (Campbell, Bownas, Peterson, and Dunnette, 1974), partly because the people who report the data know that the data will be used to evaluate them and their group. Consequently, performance data submitted to management information systems are almost always to some extent manipulated to conform to the expectations of management.

Another psychological consideration in human performance measurement is the effect of volunteerism. Since it is rarely feasible to measure the entire population, a sample must be taken. Unless one has complete power over drawing the sample (which is an unusual case indeed), the composition of the sample will be affected by the propensity of people to volunteer. Volunteerism can create a seriously biased sample because people who agree to participate frequently differ from those who do not (Rosenthal and Rosnow, 1974). The extent to which the volunteerism factor creates a bias in the results varies from situation to situation and, unfortunately, cannot easily be predicted in advance. Research

on volunteering for psychological experiments, for example, has revealed that the type of experiment (Martin and Marcuse, 1958), the alternatives to participating (Blake, Berkowitz, Bellamy, and Mouton, 1956), and the reaction of others to the request (Blake and Rosenbaum, 1955) are all important situational factors related to volunteering. The advent of the Privacy Act has made volunteerism an even more salient factor: when there is no legal basis on which to force participation, the voluntary nature of the information request must be clearly addressed in the Privacy Act Statement.

Another factor likely to affect the representativeness of the sample is the physical situation in which the measures are taken. Situations can be classified on a continuum from artificial to real-world. An intermediary position on this continuum is measuring people while they are in a training situation or in a simulation of their work environment. People's attitudes have a great deal to do with the difference in performance data obtained in these different situations. In general, people tend to be less threatened and more likely to volunteer, the farther removed the performance situation is from their operational work situation. Also, it is easy to control measures taken in an artificial situation, creating high internal validity, but frequently such measures reflect little about performance in the real world where people contend with real-world pressures and real-world distractions. Therefore, artificial situations usually have low external validity. Choice of the situation in which to take performance measures should therefore be dictated by one's purpose. If one is interested in knowing people's potential for optimal performance and the factors which facilitate or degrade it, then an artificial, laboratory situation would be appropriate. If, however, one is interested in understanding people's everyday work behavior, then measuring them in their operational work setting is preferable.

Up until this point this paper has dealt with the problems associated with doing performance measurement on humans, including the possibility of degrading performance by negative impact on people's feelings and attitudes. On the other side of the coin is the possibility of enhancing human performance through positive impact on people's feelings and attitudes. Such a possibility was probably first noted with the discovery of the Hawthorne Effect, which is an increase in performance as a result of employees being aware of their being studied. In other words, it is possible to use the process of performance measurement as a motivating device. Douglas McGregor's (1960) ideas had substantial impact here. He believed that if feedback about successful task accomplishment were given to employees, this would enhance their self-esteem and increase their motivation. Prior to McGregor's suggestion, the most commonly held theory of organizational behavior was that motivation causes good performance. However, since McGregor's suggestion, empirical evidence has been accumulating that the reverse is more often the case; that is, good performance causes motivation. Following his own rationale, McGregor developed the concept of Management by Objectives (MBO), which has the following advantages peculiarly suited to performance measurement of humans: (1) discussion is limited to output and does not deal with the employee's personality or other traits, (2) discussion is directed toward what can be accomplished in the future and does not dwell on what was not accomplished in the past, and (3) the system is designed to provide the employee (and the supervisor) with feedback on task accomplishment.

Another outgrowth of McGregor's work was the concept of the Scanlon Plan, which is a scheme to equitably share profits which result from increases in employee

productivity. In MBO, the reinforcers are primarily intrinsic, while in a Scanlon Plan they are primarily extrinsic. Accurate performance measurement is an essential component since one must be able to assess increases in productivity and to know what part of the organization is responsible for them. There are presently more than 100 Scanlon Plans in operation around the U. S., and there seems to be a recent resurgence of interest in their use (Tracy, 1977).

The use of feedback from performance measurement to enhance motivation and productivity is beginning to receive increased attention with the recent advent of a behavior modification movement within organizational psychology ("At Emery Air Freight: Positive Reinforcement Boosts Performance," 1973; Luthans and Kreitner, 1975; Nord, 1969). This movement is based on Skinnerian principles of operant conditioning. Also, it utilizes positive rather than negative reinforcement. While external reinforcement such as praise from the supervisor is used at the outset, the crux of the program is to restructure jobs so that immediate feedback is provided to the employee directly from job performance itself. Eventually, the feedback component becomes the primary motivator, especially as the employee becomes skilled at the job. Some impressive increases in productivity and decreases in absenteeism have ensued as a result of such programs (Nord, 1970), and the popularity of this approach is increasing.

PERFORMANCE MEASUREMENT OF HUMANS BY HUMANS

The significance of humans measuring the performance of other humans lies in the fact that it is a social situation. Consequently, interpersonal dynamics come into play. Again, it is useful to distinguish in terms of the purpose of the performance measurement: research, feedback, or appraisal.

In the case of research, the social interaction aspects are usually at a minimum. Ordinarily, the researcher has no previous acquaintance with the subjects, is from an outside organization, and will have no future social relationship with the subjects. At the other end of the continuum, in performance appraisal, the person doing the measuring is usually the subject's immediate supervisor. Therefore, a permanent, intimate social relationship exists between the two people involved. Since the supervisor is dependent upon the subordinate to do the required work, supervisors are understandably reluctant to say or do anything during an appraisal to alienate the subordinate. Hence the ubiquitous "vanishing performance appraisal syndrome" (Porter, Lawler, and Hackman, 1975) in which, when supervisors are asked, they report appraisal, feedback, and counseling sessions with subordinates, and when their subordinates are asked, they report having received no such sessions. Even when a formal appraisal session is required, it is frequently perfunctory and brief due to both parties' discomfort in the situation.

Volunteerism, which was addressed earlier, is also affected by the relationship between the person doing the measuring and the person who is considering volunteering. Since a personal, face-to-face relationship is not ordinarily established until after the individual has already volunteered, the individual's decision to volunteer is usually based on attitudes toward the institution or group doing the measuring. One major factor is whether the measuring group is internal or external to the organization. There are a number of issues pertaining to whether it is preferable to use an internal or external group (Huse, 1975). An internal group may be perceived as more threatening than an external group. On the other

hand, an internal group may be more well-received due to their insiders' understanding of the work conditions and problems affecting job performance. Volunteerism can also be understood in terms of social exchange theory. This theory characterizes all relationships in terms of what is exchanged between the parties. In the case of performance measurement, the subject exchanges cooperation for whatever the measurer has to offer, including approval, payment, satisfaction of an organizational requirement, etc. In order for an exchange to take place (i.e., true cooperation on the part of both sides), both parties must feel that they have obtained something which is roughly of comparable value to what they have given. Another factor for the measurer to keep in mind is that of equity; namely, that people expect to be rewarded equally to their peers for equivalent amounts of cooperation.

A very important characteristic of situations in which people are doing the measuring is that they set up measurement situations according to their expectations about performance. Here again, McGregor's ideas have been influential. McGregor was one of the first to point out that organizations are designed not around what subordinates are actually like but what managers think subordinates are like. Consequently, employee potential for performance can be severely restricted by managers who design jobs with the attitude that employees have relatively low ability and/or motivation. Similarly, performance in experiments can be constrained due to the experimenter's assumptions about the subjects' performance potential. Not only can performance be physically constrained by job or experimental design, it can also be psychologically constrained because the supervisor/experimenter implicitly communicates expectations to the individual regarding that person's performance potential. In true self-fulfilling prophecy style, the individual is likely to act accordingly.

Another factor associated with situations in which humans are doing the measuring is that humans invariably interpret the data, while mechanical recording devices do not. The extent to which the data themselves are affected is a direct function of how subjective the measures are. Even after the measures are taken, however, an interpretation occurs to explain the "why" of performance. According to expectancy theory (Vroom, 1964), for example, performance is a multiplicative function of motivation and ability. If performance is low, it could be a result of low motivation, ability, or both. The measurer is likely to make a judgment as to which are the contributing factors, and this judgment is again likely to be implicitly communicated to the employee, which, in turn, can affect performance. This judgment is partly formed on the basis of past knowledge of the employee or similar employees and partly on the basis of attribution. Attribution is the process of inferring the reason for the behavior of others on the basis of the reasons for one's own behavior in similar situations. For example, if you find a certain task boring, you are likely to assume that others find it boring also. Therefore, if someone does the task poorly, you are likely to assume that it is due to a lack of motivation rather than a lack of ability. The more ambiguous the performance (e.g., leadership), the more the attribution process comes into play.

CONCLUSION

It is approaching 20 years time since Douglas McGregor wrote The Human Side of Enterprise (1960). While his ideas had a profound impact at the time of publication, rather than fading in influence as have so many others, his ideas have

become progressively more influential. The importance of attending to the "human side" of performance measurement is but one example.

Measuring the performance of humans certainly presents a number of difficulties not encountered in the measurement of nonhumans. Those who measure human performance usually focus on the existence of these difficulties, ignoring the bright side of human performance measurement. The bright side is that there are immense possibilities for improving human performance through the measurement process. Such possibilities are limited primarily by one's assumptions about other people and by one's imagination.

REFERENCES

At Emery Air Freight: Positive Reinforcement Boosts Performance. Organizational Dynamics, Winter 1973.

Beach, D. S. Personnel: The management of people at work. New York: MacMillan, 1975.

Beals, R. L. Politics of social research: An inquiry into the ethics and responsibilities of social scientists. Chicago: Aldine, 1960.

Blake, R. R., Berkowitz, H., Bellamy, R. Q., and Mouton, J. S. Volunteering as an avoidance act. Journal of Abnormal and Social Psychology, 1965, 53, 154-156.

Blake, R. R. and Rosenbaum, M. Volunteering as a function of field structure. Journal of Abnormal and Social Psychology, 1955, 50, 193-196.

Borman, W. C. and Dunnette, M. D. Selection of components to comprise a Naval Personnel Status Index (NPSI) and a strategy for investigating their realistic importance. Report to the Office of Naval Research, 1974.

Bryant, E. C. and Hansen, M. H. Invasion of privacy and surveys: A growing dilemma. In W. H. Sinaiko and L. A. Broedling (Eds.), Perspectives on attitude assessment: Surveys and their alternatives. Champaign, IL: Pendleton, 1976.

Campbell, J. P., Bownas, D. A., Peterson, N. G., and Dunnette, M. D. The measurement of organizational effectiveness: A review of relevant research and opinion (NPRDC Tech. Rep. 75-1). San Diego: Navy Personnel Research and Development Center, 1974 (NTIS AD No. 786 462).

Edsall, G. A positive approach to the problem of human experimentation. Daedalus, Spring, 1969, 463-479.

Feldman, R. E. Experimental and quasi-experimental field techniques: The protection of subjects. In W. H. Sinaiko and L. A. Broedling (Eds.), Perspectives on attitude assessment: Surveys and their alternatives. Champaign, IL: Pendleton, 1976.

Flamholtz, E. Human resource accounting. Encino, CA: Dickenson, 1974.

Huse, E. F. Organization development and change. St. Paul, MN: West, 1975.

Kearney, W. J. The value of behaviorally based performance appraisals. Business Horizons, June 1976, 75-83.

Luthans, F. and Kreitner, R. Organizational behavior modification. Glenview, IL: Scott, Foresman, and Company, 1975.

Martin, R. M. and Marcuse, F. L. Characteristics of volunteers and nonvolunteers in psychological experimentation. Journal of Consulting Psychology, 1958, 22, 475-479.

McGregor, D. The human side of enterprise. New York: McGraw-Hill, 1960.

Meyer, H. H., Kay, E., and French, J. R. P. Jr. Split roles in performance appraisal. Harvard Business Review, January-February 1965, 43, 123-129.

Moore, O. K. and Anderson, A. R. Some puzzling aspects of social interaction. In J. H. Criswell, H. Solomon, and P. Suppes (Eds.). Mathematical methods in small group processes. Stanford, CA: Stanford University Press, 1962.

Mosel, J. N. How to feed back performance results to trainees. In E. A. Fleishman (Ed.), Studies in personnel and industrial psychology. Homewood, IL: Dorsey, 1961.

Mustafa, H. Performance rating revisited. Civil Service Journal, April-June 1969, 28-31.

Nord, W. R. Beyond the teaching machine: The neglected area of operant conditioning in the theory and practice of management. Organizational Behavior and Human Performance, 1969, 4, 375-401.

Nord, W. R. Improving attendance through rewards. Personnel Administration, November-December 1970, 37-41.

Parsons, T. Research with human subjects and the "professional complex." Daedalus, Spring 1969, 325-360.

Porter, L. W., Lawler, E. E., and Hackman, J. R. Behavior in organizations. New York: McGraw-Hill, 1975.

Rogers, R. T. Performance appraisals: Why don't they work better? GAO Review, Fall 1975, 73-81.

Rosenthal, R. and Rosnow, R. L. (Eds.) Artifact in behavioral research. New York: Academic Press, 1969.

Rosenthal, R. and Rosnow, R. L. The volunteer subject. New York: Wiley, 1974.

Sharf, J. Second-guessing the legal fair employment system. APA Monitor, April 1977, p. 6.

Supervisor's notes OK. The First Line, June-July 1976.

Tracy, H. Scanlon Plans: Leading edge in labor-management cooperation. World of Work Report, 1977, 2, 25; 32-34.

Vroom, V. H. Work and Motivation. New York: Wiley, 1964.

Webb, E., Campbell, D. T., Schwartz, R. D., and Sechrest, L. Unobtrusive measures: Nonreactive measures in the social sciences. Chicago: Rand McNally, 1966.

ABOUT THE AUTHOR

Dr. Laurie A. Broedling is a research psychologist with the Attitude and Motivation Directorate of the Navy Personnel Research and Development Center. She is currently responsible for directing research on work attitudes and motivation and their relationship to job productivity and organizational effectiveness. She received a Ph.D. from George Washington University in 1973 in organizational psychology and has over twenty-five publications in the areas of attitude assessment, personnel surveys, personnel work motivation and performance, and organizational development.

THE STRATEGY OF PERFORMANCE MEASUREMENT

David Meister
Navy Personnel Research and Development Center
San Diego, California

ABSTRACT

Human performance measurement strategy is conceptualized as a series of questions the investigator should ask about 11 variables inherent in measurement. These variables are discussed in terms of their impact on measurement efficiency.

INTRODUCTION

It is easier to say what a strategy of performance measurement is not than what it is. It is definitely not a novel experimental design or statistical technique. It is not a formal step-by-step procedure. It is, however, a logical process, requiring analysis of the variables inherent in performance measurement and answers to questions posed by these variables. The purpose of this paper is to describe these variables and their implications for measurement.

Before beginning this discussion, however, it is necessary to define our concept of performance measurement. This may differ from views of other measurement specialists because of its pragmatism and its orientation around the concept of the man-machine system.

Without going into any detail about man-machine system assumptions (the author has expressed his views on this topic in his recent book, Meister, 1976), it is necessary to make the following statements. The measurement described in this paper is focussed on human accomplishment of a total task or job performed in a work environment. The totality of performance measurement, as described in the behavioral literature, obviously encompasses much more than this. Laboratory studies of part-tasks or functions such as tests of visual discrimination or simple motor processes like bar-pressing, are outside of our ken. Likewise, the use of written tests, interviews and rating scales to determine personnel capability, i.e., aptitudes and knowledges, is not of interest here because aptitudes and knowledges, although necessary for the accomplishment of tasks, are not isomorphic with these tasks. Non-performance methods can and must be used in a performance context to help explain variations in performance, but our concern is with the performance itself.

Performance measurement is viewed as an effort to determine whether the required, desired or anticipated performance of system personnel--those who operate and maintain a system--satisfies the requirements of the system and/or the anticipations of the system developer or user. Every activity organized around and directed at achieving a specified purpose can be considered a system. Thus, an "A," "B" or "C" school (and, for that matter, the total Navy training organization) is a system; a ship is a system; an aircraft is a system; the Combat Information Center (CIC) aboard ship is a system--albeit at a lower level than that of the ship of which it is a part.

PRECEDING PAGE BLANK-NOT FILMED

The size of the unit whose personnel performance is being measured is not at issue here; it may vary from the single operator/console combination to an entire ship's complement or, for that matter, the Navy's entire training establishment. Obviously the size of the unit has serious implications for measurement (see variable 2 below), but the significance of the system concept for performance measurement is that the purpose for which a system was developed implies performance requirements; the need to measure derives from the need to determine whether system performance in fact satisfies these requirements. The personnel performance being measured is that of humans working in the system to accomplish the system purpose. The goal of the measurement is performance evaluation. The system produces an output or accomplishes a mission; that output or accomplishment can be used as a measure of the efficiency of the total system and as a means of determining the contribution of system personnel to system efficiency. This framework differentiates our concerns from those of researchers who collect data on personnel performing in non-task-oriented groups.

Moreover, we are not concerned in this paper with measurement specifically and solely for research purposes (although research information may be gathered incidental to our main goal). Performance measurement for research purposes may or may not utilize a system context (although it is our feeling that it should); but it is not evaluative because it is not concerned with a system purpose. Nor are we concerned with the gathering of normative (e.g., census) data, although such data inevitably fall out of the measurement process.

The systems to which we refer are military systems. There is no reason, however, why the measurement strategy described would not be applicable to non-military systems, such as commercial/industrial ones--automobiles or commercial aircraft --or governmental social-benefit systems like hospitals, universities or water distribution networks. The latter also fit our system definition and are developed to accomplish a purpose.

The major difference between military and non-military systems with regard to measurement is that the purposes for which non-military systems were developed do not impose as stringent requirements on their personnel. In addition, non-military systems presumably create a benefit for their clients who will not make use of the systems unless their satisfaction is sufficiently enhanced. From a performance measurement standpoint the purpose of non-military systems is to achieve a desired (rather than a required) state of affairs. Moreover, that purpose must include the anticipated benefit to clients.

What is being measured is in all cases personnel performance. However, personnel performance may not be the ultimate purpose of the measurement. In many cases personnel performance is measured to evaluate the adequacy of a system, a product or a procedure because only in this way can a satisfactory test of these be made. Only when one is measuring to determine personnel capability (can individuals do their jobs?) is personnel performance of direct interest to the tester.

At first glance therefore the performance measurement paradigm required by the system orientation is quite simple. The question to be answered is: does the performance of the system (and more particularly of its personnel, the individual operator, team or groups of teams) satisfy a (system) requirement levied on them? The question implies a comparison between (1) actual and (2) required

or desired performance; either the subject of the comparison satisfies the requirement or it does not. (Although see variable 9, Experimental Design, below.)

Such a straightforward comparison is of course overly simplistic. The variables implicit in the measurement process make the comparison in many cases very complex and difficult. To describe a performance measurement strategy requires therefore an examination of these variables which are listed below.

	Variable	Short Title
1.	Subject of the performance measurement.	Subject
2.	Number of personnel whose performance is to be measured.	Number
3.	Measurement environment.	Environment
4.	Purpose of the measurement.	Purpose
5.	Measurement questions to be answered.	Questions
6.	Performance criteria.	Criteria
7.	Performance measures.	Measures
8.	Measurement devices.	Devices
9.	Experimental design of the measurement.	Design
10.	Relevancy	Relevancy
11.	Characteristics of personnel being measured.	Sample

One can list other variables that affect the performance measurement process, such as the characteristics of the system in which personnel will perform, but the ones described here are considered to be most important.

The author may be criticized for ignoring statistical questions, but in the framework he has postulated for performance measurement elaborate statistical designs are usually unnecessary or difficult to implement. If the comparison paradigm is valid, then statistical comparison of actual performance with that required will make use of conventional statistics such as analysis of variance, Student's "t" or Chi-square. (This is not to say of course that there are no occasions when sophisticated statistical designs are required and certainly performance measurement in a research framework may demand elaborate statistical designs. Even in the latter case, however, research requirements must fit within the constraints imposed by the manner in which the system normally functions. For example, the system may impose a particular order of presenting stimulus inputs. In a purely research context the investigator may wish to counterbalance order of presentation in order to avoid possible contamination effects resulting from a fixed order, but this may not be possible in a system measurement framework without violating the inherent logic of system functioning.)

SUBJECT

The system has three major elements: the machine, the man and the system which includes both of them (as well as other elements such as the environment).

Consequently there are three sources of performance data: the equipment, the human and the subsystem or system, the last two including the interactive outputs of both the equipment and the human.

The behavioral technologist is not interested in measuring equipment performance per se (that is the engineer's responsibility), but he is concerned about differentiating between the effect of the human and that of the equipment on system output. The system output is the criterion of the success of the system and the effectiveness of any system element must therefore be judged in relation to that output. However, since the system output commingles both operator and machine inputs, in order to determine operator effectiveness in relation to that output, the former must be differentiated from machine contributions.

For example, bombing accuracy (given that one has arrived over the target) is an output largely determined by two primary factors: the resolution of the bombsight and the aiming (perceptual) accuracy of the bombadier. If actual bombing accuracy (the system output) is ±X meters (CEP), what is the relative contribution of bombsight resolution (equipment contribution) and bombadier accuracy (human contribution)? If achieved bombing accuracy is inadequate, should the bombsight mechanism be improved or should one concentrate on bombadier accuracy? Or both?

It is of course easier to talk about differentiating human from equipment contributions to system output than it is to make the distinction in actual practice. This is because of the very intimate relationship in complex systems between man and machine, particularly where perceptual and cognitive tasks are involved. Ideally one should determine machine performance capabilities (unaffected by human performance) in advance of system functioning (like boresighting a rifle to determine its inherent error), but this may be quite difficult and the investigator may have to rely on less satisfactory post hoc analyses. If the experimental design for performance testing permits (which is only infrequently) one might consider doing a multiple regression analysis to determine the percentage of variance accounted for by each factor. The differentiation must be attempted because it is inherent in the concept of performance measurement in a man-machine system framework.

Because system elements (e.g., man, machine, environment) differ from each other and from the total system output, it is necessary to ask what the performance relationship is among the system elements and between the elements and the total system. This is not merely a matter of research curiosity. It is not enough to be able to say that operator or team performance satisfies (or fails to satisfy) requirements. One must go further to determine whether satisfying or failing to satisfy requirements has any effect on the system as a whole. Because if a particular human performance does not seriously affect system output, who cares about it (except from a pure research standpoint)?

(It might be objected that the very fact that a requirement for some human performance exists demonstrates that that performance is significant to system functioning. It may be argued that all actions required by a system (so defined by being specified in a procedure) are ipso facto necessary, since otherwise why would they be specified? If necessary, then, their absence or failure to be accomplished satisfactorily would cripple the system. Realistically, however, certain actions may be required without being very important, since there is enough "slop" in most systems to permit these systems to function even when actions are inadequately performed or even not performed at all.

The determination that a human performance meets or fails to meet a certain standard is of course important in its own right. However, without knowing the effect that performance has on the system output, it is impossible to determine whether it is worthwhile to remedy a failure to meet the standard; in other words, whether one should expend resources to remedy a performance deficiency or, rather, the cause of that dificiency. In system testing one commonly finds a certain number of personnel performance inadequacies (i.e., failures to meet standard). The question then arises how this deficiency should be remedied. Since resources are always limited, only those deficiencies that are important to adequate system functioning will be allocated corrective resources. The author recalls that in the Atlas Intercontinental Ballistic Missile Operational System Test program every human performance deficiency had to be justified before corrective action was taken (Peters and Hall, 1963); and many corrective recommendations were rejected because they did not seem sufficiently important. This is still common practice in all system testing.

In the days of Atlas system testing the judgment of deficiency importance was made intuitively, on the basis of "common sense," because investigators had no way of evaluating quantitatively the impact of that deficiency on overall system functioning. And so it continues to the present day--a very unsatisfactory procedure because it means that potentially serious human performance deficiencies may receive inadequate attention if conservative project managers or operational personnel cannot be convinced that a behavioral deficiency is really important.

Determination of the human contribution to system performance means that the behavioral technologist must correlate human performance variability with system output variability. As a purely hypothetical example, if one were to plot the range at which initial ASW sonar detection is made against the accuracy of depth charge attacks, a high correlation might be found. This would suggest that detection range has a significant effect on system performance and that efforts be made to improve that detection range, either by providing more training to sonarmen or improving sonar resolution.

There is of course always the danger that little or no relationship may be found between the human and the system output. Although the measurement specialist has a tendency to believe that all human outputs are very important, his point of view may be biased. Alternatively, the system output measure selected (and in any complex system there will be more than one) may be inappropriate or perhaps too molar to permit the relationship with human performance to manifest itself strongly. Again, it is possible that a relationship between detection range and depth charge accuracy actually exists, but is "buffered" by intervening system processes. By buffering is meant that the inadequacies of intervening system processes (e.g., plotting target position, delays in firing) may combine to reduce the direct effect of initial detection range. All of these possibilities may make it difficult to secure a significant direct relationship between a particular human performance and the system output. Nevertheless, the effort must be made.

In any system therefore one's performance measurement must include individual operator or team (i.e., human outputs) and the overall subsystem or system (system outputs). This requirement has several implications:

(1) It is necessary to measure overall system output as well as the human outputs contributing to the former.

(2) It is necessary to distinguish between the relative contributions of man and machine to that system output.

(3) It is necessary to "play" the human performance contribution against the system output to determine if variations in human performance produce variations in system output.

NUMBER

This variable describes the number of personnel whose performance one is measuring. The continuum here ranges from the individual operator to the total complement of a ship, an aircraft, a regiment.

It should be obvious that it is easier to measure the performance of the individual operator than that of a team or subsystem consisting of many personnel. Easiest to handle is the individual because one can deal with him "one on one." (For that reason, most behavioral research literature deals with single operator/equipment combinations.) If his performance is being observed, a single observer will probably suffice. As the size of the team increases, the number of observers must be increased. (This is not necessarily true of instrumentation devices; the relationship between unit size and instrumentation complexity is not necessarily linear. A single equipment can record team communications as well as it does those of an individual operator.) At its extreme, as for example in the measurement of an entire ship's complement, the number of observers becomes excessive; for example, in the recent OPEVAL of the LHA-1 (TARAWA) 91 observers were required.

The logistics of performance measurement is only one, and that the least difficult, of the problems encountered as one moves from the individual to the team to the subsystem/system. Far more important is the fact that as soon as one deals with a team one must consider measurement of the interactions among team personnel. It might be supposed that all one has to do to determine whether a team meets standard is to measure the overall team output. However, the situation here is entirely parallel to the relationship between the individual operator and the system output. If team performance does not meet standard, it becomes necessary to find out why (in particular, the team member or members whose performance contributes to the inadequacy of the team output). This requires that the performance of each member of the team be measured (as well as that of the team as a whole) and also the interactions between team members, because the source of the inadequacy may be in those interactions.

Other problems arise in a team situation. It has been pointed out (Glaser, 1955) that it may be difficult to determine the boundaries of the team (in other words, those individuals and performances to include in the team being observed). This is perhaps less important than the fact that as the team becomes larger, the number of functions it performs ordinarily increases. An intensive analysis may be necessary to identify those functions that represent team interaction. Since there may be more functions to be measured than one has resources, it may become necessary to determine which functions are most important and to concentrate on these.

One sees this most clearly at the system level. At that level (e.g., the ship as a whole) certain variables which would be of minor importance for the individual operator may become tremendously important. For example, equipment maintenance effectiveness which plays a minor role for the individual operator

(unless he is specifically also a maintenance man) is a crucial parameter for the ship, since it determines (at least partially) its operational readiness and availability. These new variables may demand not only additional resources but new measurement methods. Measures appropriate at a lower (e.g., individual operator) level may not be appropriate at a higher (e.g., subsystem) level. For example, the operator's accuracy in operating a joystick may be too molecular in measuring the performance of a large command/control subsystem like NTDS.

From a strategy standpoint therefore it becomes necessary to examine the personnel whose performance is being measured in terms of their number and level in the system. It requires also a consideration of the variables and functions pertinent to and the measurement methods appropriate to that level. As the size of the measurement unit becomes larger and extends to the subsystem/system level, the problems involved in specifying what is to be measured and how become increasingly burdensome. Nevertheless, it is not feasible to "slice off" a performance segment for measurement simply for the investigator's convenience. Ideally one should measure concurrently at all levels (the system, subsystem, team and individual) and interrelate the results. Practically this may be impossible because of limited resources; and the investigator's task then becomes one of selecting the level or levels at which measurement will maximize the conclusions he can derive.

ENVIRONMENT

The measurement context must also be considered in developing a strategy for a particular type of performance. This is because context often imposes severe limitations on what one can do in performance measurement.

The places where human performance can be measured can be categorized in terms of the usual operational/non-operational or field vs. non-field dichotomy: the laboratory, the school and the simulation facility which fall into the non-field context, and the test range and the area of operations which fall into the field category.

All categorizations are arbitrary to some extent. Even within the operational environment there are gradations. For example, a missile test range like Vandenberg Air Force Base can be considered operational, but obviously it is less operational than an actual hardened missile site. For a weapon system actual combat is the only true operational environment (because it is the context in which the system is ultimately supposed to be used), but it is impossible to measure in actual combat, although an attempt is made to measure in "war games."

The factor differentiating field from non-field is the degree of fidelity to the operational environment. On that basis it may appear as if a highly realistic simulator (which may be found in both laboratories and schools) would closely approximate operational functioning. However, we define operational fidelity in terms not only of hardware similarity but also in terms of the range of factors that affect actual system functioning. Two operational factors which cannot be ignored are chance (which increases variability in inputs to the system) and noise (i.e., factors outside the system that enter and impact on system functioning). An aircraft simulator may be highly realistic but only as it pertains to hardware functions; it usually does not take into account additional factors such as a potential enemy, level of maintenance, or interaction with other combat units. From a performance standpoint the most elaborate simulator usually requires only part of the operational performance one might wish to measure, because it describes only a subsystem of the total actual system.

Actually, an aircraft simulator is a poor example; the discrepancy between simulation and operational performance is greatest when one deals with large systems like a ship or CIC. It has been pointed out (Meister, 1977) that the attempt to measure human performance in very large systems practically requires measurement in an operational environment and that for such systems traditional measurement paradigms may be very difficult to apply. Thus, if one wishes to evaluate the performance of an entire ship's complement, it is impossible to attempt this in some sort of simulator. This is not only because the amount of hardware would be prohibitive but also because it would be very difficult to simulate certain critical aspects of the large system such as its maintenance. Moreover, classic hypothesis-testing paradigms such as those that involve experimental and control groups are very difficult to implement in that environment.

There are several reasons for emphasizing the operational environment in this discussion. First and foremost is the fact that actual operations serve as the criterion to which we wish to generalize our performance measurement. If we measure for operational readiness it is because we wish to predict how personnel will perform in combat; if we evaluate new system adequacy it is because we wish to determine how that system will perform operationally. If we measure trainee performance it is because we wish to know how the student will do in operations (although this last statement must be qualified, because the school situation is often used as its own criterion, e.g., graduation as a criterion of school success).

Beyond this (1) many of the systems the performance of whose personnel one wishes to measure can be measured only in the operational environment; (2) measurement in the operational environment presents special problems, some of which have been discussed in another paper (Clarkin, 1977). On the other hand, the operational environment more readily permits the measurement of a higher order system output which is much more difficult to assess in a non-operational environment where only parts of the total system can be simulated. In the operational environment the greatest problem is of course lack of control over subject performance. Non-field (e.g., laboratory) contexts are more attractive to researchers because these permit greater control. However, they provide this opportunity only by deliberately eliminating many of the factors that make the operational environment meaningful. On the other hand, for research purposes or to determine whether personnel are learning adequately (the school situation) the non-operational environment may be more satisfactory.

The author's viewpoint is that <u>where possible</u> performance measurement should be accomplished in the operational environment. This is particularly the case when one is attempting to measure human performance in large systems, because such systems cannot be divorced from that environment. The operational environment is also necessary if the tasks being measured are complex and interact with their environment. If the tasks to be measured are unlikely to be influenced significantly by operational factors (and one can perhaps determine this by prior observation in the operational environment), then a non-operational environment is reasonable. Such tasks are likely to be relatively discrete (i.e., without numerous personnel interactions), for example, typing performance, which is likely to be unaffected by the operational environment except in extremely bad weather.

Where measurement in the operational situation is not possible--and often there are reasons why this can't be done--the investigator should include in his non-operational setting as many operational factors as possible. This requires of him a very deliberate effort to analyze the operational environment of the performance to be measured.

The strategy of performance measurement requires the investigator to decide in what environment he should conduct his measurement. He must first ask himself whether he has any choice; what are the alternatives available to him? The strategy also requires him to examine the factors that act on his subject in the operational environment, to decide how important these are for his measurement purpose; how to measure these if he opts for the operational environment; or, alternatively, how to incorporate these factors into his laboratory situation if he chooses that route. It is possible for him to make his laboratory replica more like the operational environment by deliberately relaxing the amount of control he has over stimulus inputs and by enriching the laboratory with features he has observed to operate in the operational environment.

PURPOSE

The purposes of human performance measurement have been implied previously, but not expressly described. It is important for the investigator to consider the specific purpose of his measurement because that purpose will partially determine his measurement environment and methods.

These purposes are:

(1) Measurement of the operational readiness of personnel to perform system tasks;

(2) Evaluation of systems, products, and/or procedures to determine if these satisfy development and operational requirements;

(3) Determination of personnel capabilities (do personnel possess required skills?) and training achievement (do personnel meet course graduation standards?);

(4) Research to secure normative data and personnel performance information about the conditions that affect system output.

In all cases it is the human's ability to utilize a system or product and to perform job-related tasks that is the focus of the measurement.

The individual purposes are not mutually exclusive; they are not, however, usually combined in a single measurement situation (except that research data may be secured along with each of the three preceding purposes). By implication the purposes are linked; for example, measurement to determine operational readiness may produce data about personnel capabilities, insofar as the latter are needed if the system is to be operationally ready.

As we have seen, the measurement of operational readiness in almost all cases requires the collection of data in the operational environment with personnel performing required tasks in as close to combat mode as possible. The determination of operational readiness by means of paper and pencil tests or ratings of personnel capability is an inadequate means of making this determination.

The evaluation of systems, products and procedures may take place in special test facilities, e.g., test firing ranges, as well as in operational areas, but every effort should be made to include at least the most important operational conditions in such facilities. When such evaluations are conducted during system development, the earlier in development that such measurements are made, the less possible it is to simulate operational conditions.

Determination of personnel capabilities usually requires some control over the conditions of stimulus presentation; hence a simulation or school situation is to be preferred. Measurement of training achievement usually takes place in the school environment. Since personnel capability and training achievement are directly tied to the individual, the performance recorded is usually that of a single individual only (except in the case of team-specific training). Often the total system context is not required for these measurements and only those equipment elements directly related to the skills and knowledges being measured will be utilized, this is because skills and knowledges are not directly translated into system performance and are meaningful for system performance only in the form of job-required tasks.

QUESTIONS

The essential question which performance measurement in the system framework seeks to answer is: does personnel performance satisfy system requirements and to what extent? The only exception is the research goal which, as indicated previously, is outside the purview of this discussion.

The answer to this question is insufficient, however, to satisfy all measurement goals. What happens if personnel performance fails to satisfy system requirements completely? It is then necessary to ask further questions; specifically:

(1) Who (the individual, the team or subsystem) is responsible for or contributes most to the performance inadequacy?

(2) What is the cause of the inadequate performance?

In addition, the investigator should ask what basic information (leading to further knowledge of how personnel perform) can the measurement provide?

If all personnel performances are satisfactory, questions (1) and (2) above need not be asked. However, it is rare to find any system without personnel-related inadequacies. In a ship whose operational readiness is certified as adequate, certain subsystems may still be marginal or even deficient; in the evaluation of a new system or product certain aspects of that system or product may be deficient, even though the system or product as a whole passes required tests. In a training evaluation (school) situation some personnel fail. It therefore becomes necessary to identify the deficient measurement unit so that it can be improved; to remedy the deficiency its source must be known.

To identify the subject responsible for inadequate performance one must be able to differentiate among the contributions individual subsystems, teams or operators make to system output. (The assumption is made that usually more than one individual is involved in system operations.) In some cases one can do this analytically; since certain individuals are "key" personnel (occupying a central position relative to system outputs) it is possible to relate serious inadequacies to such personnel because logically their role makes them responsible (at least

for the quality control of their subordinates' performance). Where it is not possible to deduce the inadequate subject, it becomes necessary to trace the contribution of each job position to the system output. This may also require the investigator to disentangle man from machine inputs, a problem discussed previously and one which is difficult when the investigator cannot experimentally manipulate operations.

The determination of the deficient subject is only one step in the determination of the cause of ineffective performance. To determine who is responsible is not the point of the analysis, although knowing who is essential to finding out what; the point is to determine the cause of the deficiency so that the underlying conditions can be altered. Since the ultimate rationale of performance measurement is to optimize the system, this means remedying all significant deficiencies --or at least as many as one has resources for.

The necessity for determining who and what are responsible for deficient performance has implications for the methodology employed. Under ideal circumstances the question, do subjects perform to standard, can be answered by completely objective means; but questions relating to causal factors demand more subjective methods. In the author's experience very few systems are so designed that a performance deficiency immediately points to its cause; there is an intermediate analytic step, which must be supplied by the performer himself or by an observer. The performer can often supply vital information relative to the problem; the observer (hopefully an expert in system operations) can discern something through his observation that reveals the cause of the deficiency. This may require use of a wide spectrum of subjective techniques such as the interview, the questionnaire, rating scales, observational checklists, etc. These subjective techniques are generally denigrated by measurement professionals but in actual measurement they are invaluable.

Unfortunately, outside of their use in research these instruments are not systematically developed and validated so that in addition to their inherent subjectivity we face potential problems of invalidity and unreliability. Nevertheless, one essential feature of any performance measurement strategy (including a strategy for research) is never to complete a series of measurements without cross-examining the performer concerning what he has done, the problems he has encountered, the reasons (he supposes) for those problems, the ways he has found to overcome them.

Although the goal of our performance measurement is not research, occasionally situations inherent in system characteristics or functioning permit the collection of data that have a research as well as an operational significance. For example, if the system (and its personnel) must perform under both day and night conditions, the opportunity exists to collect data under the two conditions, data which, when compared, will enable us to say something about the effect of these conditions on performance. Such opportunities should not be ignored. The experimental design for such comparisons is relatively simple, since the data must be collected in conformance with system constraints.

In general, the more information one collects about how system personnel perform, the better, since there is a distressing lack of knowledge about how personnel function, particularly in complex macro-systems. Certainly one cannot rely on traditional sources of behavioral research data, because few researchers are concerned with the system or can deal with it in a laboratory context (Meister, 1975). Any system performance measurement therefore provides an opportunity to collect data useful for research purposes, but unfortunately those who

measure performance in the military setting usually ignore available research possibilities. It is in fact remarkable that so little use for performance research is made of the vast military population in their daily routine.

CRITERIA

The essence of performance measurement as it has been described here is the criterion. Without it no meaningful measurement is possible; without it, one can collect descriptive performance data (to answer the question, what are personnel doing?), but the meaning of those data--whether personnel are performing adequately or not--cannot be determined.

There are three distinct types of performance criteria: those describing the functioning of the system; those describing how missions are performed; and those describing how personnel respond. Only the last is of interest to us, but the fact that different criteria are available means that it is necessary to differentiate among them. System-descriptive criteria include such aspects as reliability, maintainability, vulnerability, cost of operation and effectiveness. Only the last (effectiveness) requires differentiation between human and machine inputs, since effectiveness criteria mix equipment and personnel elements. This is true also of mission-descriptive criteria which include output quality and accuracy, reaction time, queues and delays.

Each of the preceding includes personnel elements that must be differentiated from non-personnel elements. At the same time personnel performance criteria describing individual operator and crew responses (reaction time, accuracy, response number, speed, variability, etc.) are insufficient unless, as was pointed out previously, they are considered in relation to system and mission criteria.

"Performance criteria may act as independent or dependent variables. As independent variables (e.g., the requirement to produce N units) they impose on the operator a requirement that serves as a forcing function for operator/system performance. As dependent variables they describe operator/system performance and can be used to measure that performance." (Meister, 1976, p. 13)

The criterion therefore implies a demand imposed by the system on its personnel. Personnel must or should do something to accomplish some goal. The investigator must differentiate criterion-referenced data (which imply a standard of comparison) from purely normative data (people do thus and so).

In evaluative performance measurement the criterion implies a standard of performance acceptable to the system mission. In measurement research criteria are also necessary (as dependent variables which describe the effects of independent variables) but often they do not imply or require standards. Although it is possible in research to have criteria without standards, in evaluative measurement a criterion is meaningless without a standard because it does not provide a measure of determining whether personnel are performing well or poorly. Thus, for example, the number of messages decoded is a criterion which can be used in research on intelligence systems; but the evaluation of intelligence personnel makes it necessary to specify in advance of measurement that they decode N messages per hour. Henceforth when we use the term "criterion" it must be understood to imply a standard.

It is not enough, however, to have a criterion; the criterion must be precise or it cannot serve its purpose. To be precise it must in most cases be quantitative as well. A criterion such as one which is often found in system procurement descriptions, "the system shall be so designed that personnel perform their duties with minimum difficulty," is meaningless because it is undefined; or rather it can be defined only in terms peculiar to the evaluator. With undefined criteria one must rely on the evaluator's ability to translate the criterion into concrete terms; and without those terms being specified in writing it is almost impossible to communicate their meaning to others.

The specification of precise quantitative criteria presents a number of difficulties which the investigator can overcome by persistence and good humor.

For example, it is extremely difficult to persuade military personnel (even those having the greatest familiarity with a functioning system) to provide precise criteria of how that system functions. The usual response is, "it all depends." This may reflect the feeling that the performance depends on so many interactive factors that it cannot be specified (although presumably it can be recognized by experts). On the other hand, military personnel may fear that if they specify precise criteria, their performance will be judged too stringently.

In systems under development one often finds that although hardware performance criteria are specified in explicit terms, there are few or no references to personnel criteria. In part this reflects a wide-spread impression among system developers who lack behavioral background that personnel performance either does not matter to system outputs or is too variable to be described. Of course, some systems may require so many contingent responses that it is difficult to supply standards for every contingency. However, even in such systems it should be possible to supply precise criteria for the major outputs required of system personnel.

Not all criteria are equally relevant and valuable for performance measurement. The level of adrenalin in the blood of subjects performing a visual vigilance task has been shown to be related to target detection (see Baker et al., 1970), but adrenalin level is not the most desirable criterion one can find to measure sonar detection because it is only indirectly performance-related. The investigator should examine the criteria he has available and select those that seem most directly related to the performance at issue. The relevance and importance of a potential criterion can be determined by asking how seriously the achievement of or failure to achieve the criterion will be for system performance.

For example, if one were to contrast false alarm rate and adrenalin level in sonar performance, which would impact more on target detection? If this impact is slight, the potential criterion is not a very satisfactory candidate. In other words, the criterion falls out of what is required of system personnel and whatever affects them strongly represents a potentially usable criterion.

If the first step in setting up a performance measurement program is to develop a plan for that program, the first step in developing the plan is to ask, what must personnel do (the criterion)?

If the answer to this question is unknown (that is, no personnel criteria have been specified), it is possible to develop criteria by using skilled operational personnel to secure the answer. (Of course, this applies only when the system

is already developed or, if under development, is similar to ones already in existence.) In the formal procedure for securing such operational judgments, the Delphi technique (Dalkey and Helmer, 1963), operational personnel are called together and required to specify quantitative criteria; the variance in these judgments is progressively reduced by successive Delphi sessions until an acceptable quantitative consensus is achieved.

What the measurement specialist is looking for is an objective quantitative criterion, e.g., an operator is expected to decode 16 messages per hour or to detect all targets at 1000 meters. With such criteria the performance described by the criterion can be observed and recorded without intervention either by observers or by the personnel whose performance is the subject of the measurement. Unfortunately, many criteria cannot be objective and quantitative. Some performances (primarily perceptual and cognitive) are inherently subjective. If, for example, the criterion is quality of decision making in a combat situation, it may not be possible to measure this with instrumentation. The cues needed to describe quality may be so tenuous that only an expert can perceive them.

There is no reason to discard criteria that cannot satisfy objective requirements. Such subjective criteria can be strengthened by consensus techniques such as Delphi. However, it is apparent that criterion precision will determine how adequately one can measure and by what methods. For qualitative criteria we must call upon the expert because only he has the requisite experience to recognize the performance involved. We can accept conclusions based on such criteria, but with a somewhat lesser level of confidence.

It is not acceptable, however, to rely on subjective, qualitative criteria when more precise, objective criteria are available. This problem may arise with inexperienced personnel. In one illustrative situation, involving the performance of infantry disembarking from a landing craft after a prolonged sea ride, a high ranking officer indicated his preference for an observational judgment (based on officer experience) of their capability to engage in combat, rather than objective measures or running (speed), climbing (agility) and firing (accuracy).

Complex systems may also have multiple performance criteria because personnel must perform a variety of functions. If so, one must measure them all; the investigator should not pick and choose (especially not post facto) even though it may be embarrassing when he secures positive (desirable) results with one criterion and negative (undesirable) results with another. The author recalls one study he performed (Meister et al., 1971) of the training effectiveness of the S2E aircraft simulator; multiple criteria suggested that the simulator trained certain functions well and others not at all. Since the aim of the study was to demonstrate trainer utility, the sponsor of the study was not overly pleased with the results.

Criteria interact with other variables, such as subject and number. As one proceeds up the ladder from individual operator to team or from subsystem to system, the nature of criteria will change. In measuring team performance, for example, one must consider member interactions, a criterion which is obviously irrelevant to single operator performance.

Our strategy of performance measurement places major emphasis on development of performance standards. If these are lacking, no evaluation measurement is possible, although one can gather normative (i.e., descriptive) data. If someone objects that even without expressed standards he can look at a set of data and determine that performance is adequate or inadequate, our rejoinder is that he does in fact have a (mental) standard, but has failed to specify it.

All potential criteria must be analyzed in terms of their eventual use as standards. Some may later be rejected as of minor importance, but it is not permissable to concentrate on a few outstanding or immediately apparent performance features and ignore the rest. This procedure may lead the investigator to overlook critical aspects of personnel/system performance for which criteria are obscure.

MEASURES

Performance measures have a direct relationship to criteria. In fact, if the specialist has difficulty finding a measure to describe the criterion he has specified, the latter is probably incorrectly described. Nonetheless, the same criterion can be described by a number of measures. For example, suppose the criterion is effectiveness of corrective maintenance. One obvious measure is downtime: the time it takes the technician to restore a malfunctioning equipment to operating condition. However, other measures are possible. For example, the number of malfunctions correctly diagnosed or the speed of malfunction diagnosis (not the same as remedying the fault). In the school situation knowledge can be used as a measure, although this is not a performance measure.

Often it is difficult to distinguish between a criterion and a measure, and some investigators confuse them. A criterion like effectiveness is relatively molar and cannot be directly defined in terms of personnel actions; a measure is quite specific and (for performance measurement use) must be described behaviorally. (Of course, non-performance measures like knowledge are not described in terms of personnel actions.) It is possible that the more molar the criterion, the more measures are available to describe it. For example, in evaluating the performance of Civil Servants, the Navy Personnel R&D Center uses a number of measures (although unfortunately not performance-oriented ones); we evaluate personnel on the basis on knowledge of the profession, procedures, specifications, etc.; ability to write, communicate, perform research, etc.; personal attributes such as responsibility, adaptability, etc. General categories of measures have been described by Smode et al., 1962, and in Meister, 1971.

In any event, in moving from the criterion to the measure it is necessary to specify the operations which reflect the criterion; these operations then become measures.

Since for almost every criterion a number of measures is available, the investigator must decide which one or ones he will use. Since each measure reflects a somewhat different aspect of the criterion, each measure may provide a slightly different result. Nevertheless, it is our recommendation that within the limits of his resources and the demands of the measurement situation, the investigator should employ all of them, even at the risk that some measures will provide him with discrepant results.

If it is necessary to select among measures, however, we have found it useful to select those which are:

(1) Objective. Ideally the measures employed should depend as little as possible on human judgment, because data collection in which the human is the measuring instrument inevitably involves considerable inaccuracy and inconsistency. As a matter of practicality and cost, however, many measures employed in performance measurement cannot be completely objective.

(2) Quantitative. Quantitative measures can be scaled and combined with other quantitative data; this is not true of qualitative data.

(3) Unobtrusive. The act of gathering data should not affect the manner in which the operator performs his tasks. All data collection agencies (instrumental as well as human) should ideally be invisible to the performer. If personnel become unduly aware of these agencies, they may perform in ways that are not representative of their routine activity.

(4) Easy to collect. Any measure whose implementation makes excessive demands (a difficult perceptual discrimination or computation) on the capability of data collectors (whether as observers or recorders of data) is likely to produce errors in the data gathering process.

(5) Require no specialized data collection techniques. There are several reasons why it is undesirable for data collection techniques to be highly specialized. Such techniques make it necessary to provide extensive training for the data collectors. More important, special data collection techniques are likely to make it impossible to use operational personnel from the population whose performance is being sampled as data collectors, because they will lack the background needed to collect the data. It is always desirable to utilize operational personnel as data collectors, first, because their familiarity with the task being evaluated may improve the precision of the data they collect; second, because operational personnel are less likely to be viewed by other operational personnel as obtrusive elements.

(6) Require no specialized instrumentation. If the performance measurement is being conducted in an operational environment, specialized instrumentation may not function well. Such instrumentation is often too delicate for the rough usage it may encounter. Also, sophisticated instrumentation will require specialists to operate and maintain it.

(7) Cost little or nothing. Cost is often the reason given by test managers or operational personnel for not wishing to conduct personnel performance tests. In most cases this is only a rationale for rejecting procedures that these managers do not understand, but obviously specialized measures may require special instrumentation and personnel, and these may indeed be costly.

Criteria for measure selection are of course ideals and in the real world of performance measurement it is often impossible to satisfy these criteria completely. The reason for listing them, however, is to provide a standard at which the investigator can aim, but we do not insist inflexibly on them.

Performance measures are utilized to answer the basic question: can personnel perform to a standard? The answer to this question, while necessary, is not sufficient. One must also ask, what is the cause of a performance deficiency (a discrepancy between actual performance and the standard). A separate set of measures must be developed to answer the question of causality. If, for example, the investigator asks a test performer, why did you do thus and so, the question "why" is a causality measure, although it is not necessarily derived from a criterion and cannot be scaled (although the responses can be categorized by content and the frequency of types of responses can be ascertained). Although there are those who will not consider the questions asked in an interview as measures, in an operative sense they are, because they provide data. In fact, in any particular measurement situation the number of objective measures derived from criteria may be relatively few; measures seeking to explain the cause of performance variability may be quite numerous. However, the latter can be used only for explanatory purposes; since they are almost always qualitative, they cannot be used to evaluate performance.

Measurement in the operational environment usually forbids highly molecular measures. Measures such as the frequency of eye movement in scanning a display usually require instrumentation which the operational environment makes very difficult to use. Moreover, only those measures which are closely related to the task being measured should be selected. If the objective of performance measurement is to determine the efficiency with which a command/control system functions, it is unlikely that measures at the level of eye movement would be considered, even though eye movement is in fact involved in scanning command/control displays. The level of measurement therefore suggests the level of the measures to be selected.

DEVICES

A device is any method used to provide data. It may be hardware; it may be a paper and pencil test; it may be an interview, questionnaire or rating scale. Observation is also a device.

It would be ideal if there were a one-to-one relationship between a performance measure and the device used to measure that performance. An objective quantitative measure (e.g., reaction time) would then call for appropriate instrumentation (a timer); a subjective measure (e.g., an operator's attitude) would call for a subjective instrument (e.g., a rating scale). The specification of a device would then be obvious and immediate.

Unfortunately the relationship is not a direct one because, first, a number of alternative devices can be used to provide data for a given measure; and, second, the constraints of the measurement situation may prevent use of the most desirable device.

Frequently the operational environment (or other factors such as cost) may require that a subjective technique be substituted for instrumentation. Obviously the subjective technique cannot be substituted unless it is a reasonable alternative. For example, the author once supervised a helicopter navigation flight test program in which a primary measure was the deviation (distance) between a specified route the aircraft had to fly and the actual route navigated (Fineberg, 1974). The original and preferable measurement concept was to use a low level radar to track the actual position of the test aircraft (the designated route was of course known). However, at the time the Army could not

provide such a radar. As an alternative trained observers flying behind and above the test aircraft were used to make visual estimates of the aircraft's deviation from the required track; because only a few routes were flown on repeated occasions, the observers achieved an accuracy of ±50 meters, which was found to be adequate (based on operational requirements) for test purposes. The point is that realistically it may not be possible to use the most desirable measurement device and one may have to fall back on a more feasible but less desirable one.

Researchers commonly feel that hardware instrumentation is to be preferred to subjective devices. Sometimes, however, the nature of the phenomenon being measured requires a subjective device. For example, in the helicopter study already referred to one of the variables was quality of flight performance. No instrumentation existing to measure quality of flying a helicopter, it was necessary to utilize a rating scale completed by the chief pilot who was also the performance observer.

A more pragmatic factor is that many project and test managers prefer subjective devices because they avoid the cost of hardware. One must also consider the complexity of the measurement situation. Professionals are tempted to utilize the most sophisticated measurement devices available even when the situation does not require them.

As a strategy therefore it is desirable that the investigator systematically consider all possible ways of recording measures and then tradeoff the variables involved in each device against the other. Pertinent considerations include, besides those criteria noted previously in the selection of measures, such factors as the length of time required to gather data with the device, its reliability and acceptability by subjects and ease of analyzing resultant data.

If one determines that a subjective device must be used, the investigator will often find it necessary to develop that device himself. Standard hardware instrumentation units can be procured off-the-shelf, or in the worst case must be developed by combining standard components, so that the investigator in most cases does not have to develop his devices "from scratch." Any subjective methodology must, however, not only be developed, but requires testing to ensure its validity. Many subjective devices are peculiar to and must therefore be developed anew for each measurement task.

It may appear superficially--and incorrectly--that it is easier to utilize subjective methods than objective devices. After all, to develop a subjective tool requires one merely (in most cases) to write something on paper. This incorrect impression derives from the fact that almost no one in performance measurement (except in research and sometimes not even then) ever systematically develops and validates a subjective performance measurement device. For example, how many times are questionnaires tested and verified? Almost never, perhaps because the questionnaire appears to be simple and directly related to the subject of the investigation. Even the ubiquitous interview, for which no preparation is usually made and which is therefore consistently abused, requires specific development, testing and validation. We may appear to be overly careful in relation to the interview, but where the performance being investigated is complex, the interview schedule becomes equally complex. More indirect methods, e.g., rating scales, attitude checklists, etc., require even more development and validation. The most difficult subjective tool to handle is observation. What is to be observed? What cues exist for recognizing the

event to be observed? How reliable is the observation from observer to observer? Where the phenomena to be observed are largely cognitive or perceptual, an extensive "front end" analysis of the task to be observed is required. Since the observation is only as valid as its observers, training must be provided to the latter and measures of their reliability (at least that) determined. The point for the measurement specialist to remember is that the subjective devices he relies on are so much more complex (conceptually speaking, that is) than objective ones that he must and should spend much time and effort to create them properly.

DESIGN

Human performance measurement in a system context permits only a limited number of experimental designs. These include:

(1) Collection of normative performance data; no comparison with a standard (single group). Sample question to be answered: How do electronic technicians perform corrective maintenance?

(2) Comparison of personnel performance with a required standard (single group). Sample questions to be answered: Is the system ready to perform operationally? Does the system satisfy requirements?

(3) Comparison of two or more alternative subject samples, system configurations, procedures, job aids, etc. Sample questions to be answered: Which training mode is superior; which technique for malfunction diagnosis is more effective?

(4) Comparison of a subject sample receiving special treatment with a control group receiving no special treatment. Sample question to be answered: Is training effective?

(5) Comparison of a subject sample before and after receiving a special treatment (pre- vs. post-test comparison). Sample question to be answered: Have personnel learned?

(6) Comparison of a subject sample receiving a special treatment in one environment with required performance in another environment (e.g., the classic transfer of training paradigm).

(7) Determination of personnel performance as a function of time, repeated stimulus inputs, etc. (as, for example, determining when performance deteriorates as a function of fatigue or workload).

Primary emphasis in this discussion has been placed on comparison of personnel performance with a standard. In the determination of individual personnel capability that standard is the task; in the determination of whether the system can satisfy system requirements the standard is the system requirement, which is some measure of mission effectiveness. Manifestly in these situations a requirement or standard must be available. Where such a requirement or standard cannot be ascertained (for example, the number of actual targets to be detected by a sonarman), one can collect normative data (how many targets do sonarmen on the average pick up?) and perhaps, considering this as a limiting measure of capability, one can transform this value into a standard which all sonarmen should achieve. From an experimental standpoint, the setting up of such

measurement situations is comparatively simple; there is only the single subject sample; the major difficulty is analytic (specifying in advance the requirement or standard to be used in evaluating the personnel performance).

The comparison of two or more alternatives is to be found most often during the development of the system but may occur also when one wishes to introduce a new procedure or job-aid into an already functioning system. The personnel acting as subjects for this comparison may be themselves, as when the two alternatives are utilized sequentially (same subjects performing under two different conditions); however, order of presentation effects must be considered here and this complicates the measurement design. Far more common when the alternatives are exposed to the operational system is to give the two treatments to two different subject samples, as, for example, trying out two training methods aboard two separate ships, e.g., CAI on one ship vs. programmed instruction manuals on another. Here one runs into the classic problem of ensuring that the two ship crews are essentially identical on all variables that could contaminate the data, e.g., Navy experience or aptitude. It is possible to make such comparisons on operational systems, but the differences between the treatments which one is willing to accept as significant must be greater than one would ordinarily require for acceptance under more controlled conditions.

In the evaluation of a training system, procedure or curriculum, comparison with a control group or a pre-, post-test comparison (or a combination of the two) is common. Neither poses any significant difficulty for inclusion in the system testing framework.

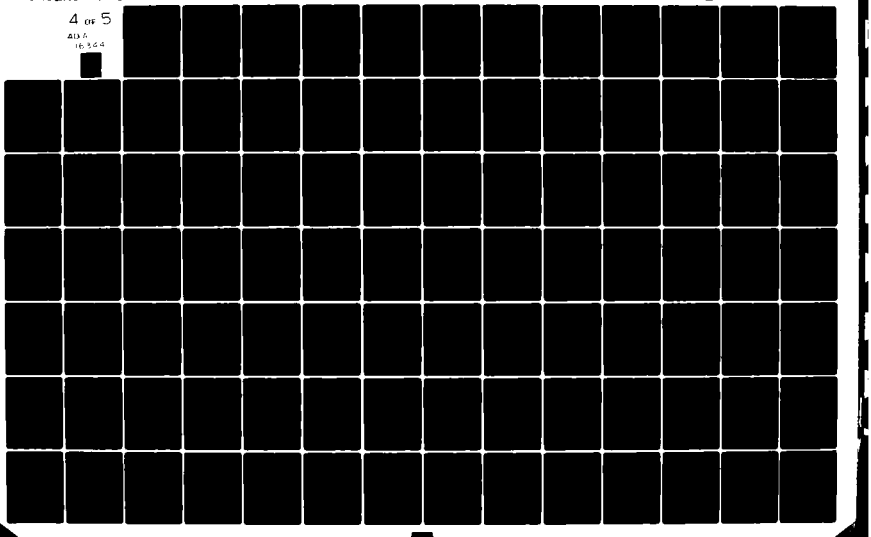
The transfer of training comparison is very difficult to implement completely. There are comparatively few problems in training and evaluating in the training environment, but the difficulties of testing the same subjects in the operational environment are very great. The problem is one, first, of tracking the trained individuals through the operational system; second, of securing permission to test their performance in that system; third, of ensuring that their operational activities are those that are relevant to their previous training environment. Suppose one wished to determine whether flight simulator training in carrier landing techniques significantly reduces the amount of overall training time required. One can measure performance at the conclusion of training and compare simulator-trained subject performance with that of personnel trained only in flight exercises; but this is only an intermediate criterion: performance aboard ship is the ultimate criterion. Having once tracked trainees to their ships, the investigator may find that some subjects are assigned to non-flight duties; and in any event the opportunity to measure carrier landing efficiency may be severely constrained.

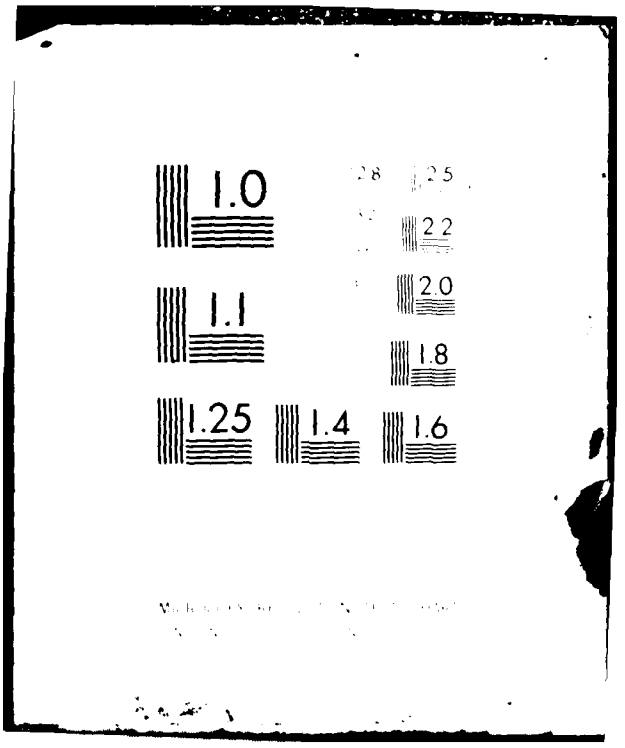
Research on the effects of workload on personnel performance is of great interest to military researchers because the combat environment with its severe stresses often leads to performance degradation (or so we suppose). The research literature on personnel performance in systems (as summarized by Parsons, 1972) has often addressed the questions: what is the progress of this degradation and at what level or stress--workload--does personnel performance break down? These questions are easy to answer in a laboratory, but very difficult to address in the operational environment, because in functioning systems the presentation of inputs is often under no one's control and rarely--except in combat--approaches the level at which one would expect breakdown. One must therefore examine the question in the laboratory or in a simulation of the operational environment.

AD-A116 344 NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 5/9
SYMPOSIUM PROCEEDINGS: PRODUCTIVITY ENHANCEMENT: PERSONNEL PERF--ETC(U)
1977 L T POPE, D MEISTER

UNCLASSIFIED NL

4 OF 5
AD A
116 344




1.0
1.1
1.25
1.4
1.6
2.2
2.0
1.8

The problem with this is, as has been indicated, the discrepancy between the operational environment and the laboratory in terms of the factors that serve as inputs to personnel.

RELEVANCY

What is being measured should be immediately apparent to the investigator. In point of fact it is necessary to distinguish between task-related performance and idiosyncratic behavior. The subject throwing switches on a control panel performs; shuffling his feet, he behaves. All performance includes behavior, but behavior is not necessarily related to the task.

In most tasks it is quite easy to make this distinction both analytically and during observation of performance. Where, however, tasks have a heavy cognitive perceptual or communications component (in other words, where some performance-related behaviors are covert) the distinction may be less easy to make. For this reason it is desirable for the investigator to extract from his system procedures the specific actions to be recorded or observed and the cues for that observation (task analysis). This is particularly important when one is relying on observation as a major measurement device.

What does one do with difficult-to-distinguish actions? Inevitably one is forced back on the use of the expert observer whose judgments are the basis of the distinction between performance and behavior. Not very satisfactory, perhaps, but the measurement specialist can at least attempt to "calibrate" his expert to remove as much as possible the subjectivity and variability of his judgments. Calibration will involve systematic training and testing.

Even more important than the distinction between performance and behavior is the relevancy of the performance being measured. We define relevancy as the similarity of the subject's performance in the measurement situation to that performance he would ordinarily manifest in actual operations.

The further removed the measurement situation is from the operational environment, the more critical the relevancy of the performance being measured becomes. If, for the military, we consider the ultimate operational situation to be performance under combat, we can never measure under true operational conditions. Even if we consider the operational situation to be, for example, normal ship steaming (Condition IV), we cannot say that we measure under fully operational conditions. The very fact that we measure (even unobtrusively) exerts a sort of Heisenberg effect on what we measure.

Relevancy must be distinguished from the concept of validity to which it is, however, closely related. Validity describes the measurements extracted from performance and asks whether these are "truthful" in the sense of describing what activities went on in that performance. Relevancy relates to the criterion performance, not to the data derived from measured performance; it asks merely whether performance in the measurement situation is the same as it is in the operational situation. Since the criterion reference for performance measurement is the operational environment (whatever the operation is), performance which cannot be related to that environment is irrelevant. Measurement may be valid in the sense that it truthfully describes the performance that occurred in the measurement situation, but even so both that performance and hence those measurements may be irrelevant. Take the classic sonarman staring intently at his PPI display in his darkened cell in the bowels of the ship. Any performance

we wish to measure must relate to this situation. If we place him in a lighted room, give him a secondary task to perform, such as adding up columns of numbers while he scans his display, then measure his target detection performance, the measures we derive may be entirely descriptive of (valid for) this laboratory situation, but completely irrelevant to his actual operating situation. For a further discussion of this point see Meister, 1977.

The operational environment includes not only functioning hardware but very complex situational conditions, the most important of which is load, which may be difficult to incorporate in a non-operational measurement. The inclusion of these conditions in the measurement makes the situation one of "worst case" because inevitably performance degrades under these conditions. Because of this such conditions may reveal embarrassing deficiencies and a level of operator/system performance lower than the investigator might wish to reveal. The author once worked on a project to evaluate an air defense surveillance system. The correct way of measuring performance in such a system is to load the system by throwing masses of "enemy" aircraft at it (simulating, for example, a heavy bomber raid) and to see how personnel cope. However, this might lead to rapid saturation of the system. Those in charge of the evaluation decided instead to expose the system to aircraft but with the following non-operational reservations: only one or two aircraft at a time; the time at which the aircraft would appear and their general heading were known in advance by operators. Under these circumstances the system was evaluated as performing quite well.

To the extent that we eliminate load factors from the measurement situation, our results become less realistic, but they look better, because personnel perform better in a non-stressful situation. This is particularly true of evaluation for operational readiness and for new systems and products; it is much less important for school-passing evaluations.

If we cannot factor these operational conditions in to the measurement situation, it is desirable to degrade measurement results by a certain amount to extrapolate to "true" operational situations. The problem is that the correct value to be used in degrading measurement data is almost never known. As a strategy ploy it is possible to ask operational personnel to rate the representativeness of the measurement situation and to supply a number which represents the extent to which the measurement situation deviates from reality. To the author's knowledge, however, this procedure has never been implemented.

SAMPLE

If one measures in the operational or school environment, it is likely that the personnel whose performance is the subject of the measurement will be reasonably representative of those to whom one wishes to generalize the measurement data. (However, where several crews perform in the operational environment, the investigator may be allowed to measure only less qualified and less desirable ones.) In the case of system/product testing, however, where the test situation merely simulates the operational environment (at a test facility, for example) and where the investigator has to select or create a subject sample with which to test, the nature of that sample may be important to the relevancy of the test results.

Fortunately there are only a few subject dimensions which are relevant to the test situation; aptitude (e.g., as measured by selection tests), experience and training. Personality characteristics may be--probably are--important but there is so little information about the relationship between personality and task performance that it is necessary to ignore this factor.

If the investigator has a choice of subject backgrounds and if these can be rank ordered in terms of relevant dimensions, should he select the "average" man (50th percentile); the least qualified (5th percentile); the most qualified (95th percentile); or a sample representative of the total continuum? There are advantages and disadvantages to each choice. A highly trained, highly experienced subject will probably give the investigator a higher level of performance and make the new system/product appear better than it may actually be. He may not, however, be representative of the total operational population. On the other hand, a less qualified subject (e.g., 5th, 50th percentile) may perform less effectively (and hence the system/product will appear less effective); but he may be more representative. Less qualified subjects may also respond in ways that reveal weaknesses in the system that need upgrading.

Ideally one would seek to have a spectrum of capability in one's subject sample, but the number of personnel available to act as subjects may be so constrained that a choice of capability may be necessary. (You can have any ten airmen--but which ten?) No definitive answer to the question is possible since the answer is likely to be determined by the level of system performance desired and is thus dependent on the individual test manager. Whatever choice is made, from a performance strategy standpoint, it is imperative that the investigator examine the capabilities of his subject population, first, to make a choice if a choice is necessary; second, even if no choice is possible, to anticipate the potential effects of his subject sample on his test results.

A STRATEGY OF PERFORMANCE MEASUREMENT

With all the variables that affect measurement, one wonders how often measurement is performed optimally. The author admits to a certain cynicism about the frequency of properly performed evaluations. In his experience personnel performance measurement as conducted by all the military services is seriously lacking, at least in two vital areas: determination of operational readiness and testing to determine whether systems, products and procedures meet operational requirements. (We need not concern ourselves with developmental testing because this is not performed to as rigid rules as operational testing.) Personnel performance measurement deficiencies in military testing arise not so much because technical expertise is lacking as much as the will to utilize that expertise. Most military testers are untrained to deal with the personnel aspects of that measurement, nor does it appear that they are overly interested in receiving assistance from those who are. The reasons for this lack of interest would take us somewhat afield from the subject of this paper.

Having examined the variables in personnel performance measurement, one must ask: what does a strategy of performance measurement consist of and how does one develop it?

This strategy derives from the variables inherent in performance measurement (the ones discussed in this paper) and therefore cannot be considered as remarkably novel. As was indicated at the start of the paper, the strategy is not a formal procedure with defined sequential steps. Rather it consists of a series

of questions which the investigator should ask about his specific measurement task. The answers to these questions will direct his actions; once answered, they must be formalized in a test planning document as described in Meister, 1966, because any strategy which is not written down is subject to misinterpretation. The major questions to be asked are:

(1) What is the purpose of the performance measurement and what measurement questions do we wish to answer? What kind of answers should the data provide?

(2) What system level and subject unit does the measurement describe? How is the subject unit defined?

(3) What criteria are available as performance standards? (If these do not exist, they must be developed by investigation and/or use of Delphi-type techniques.)

(4) What is the measurement context? How representative of operations will it be?

(5) What measures and measurement devices are available to answer questions? Which are best? Which are feasible?

(6) What characteristics should the subject sample have?

Asking (and answering) these questions will not guarantee that in any particular situation performance measurement will be optimal; but it will guarantee that the investigator can anticipate most of the measurement problems he is likely to encounter.

REFERENCES

Baker, C. H. et al. Human monitoring performance. Final Report, Contract Nonr 4120(00)-NR 196-D35, Office of Naval Research. Human Factors Research, Inc., Goleta, CA, July 1970, 7-8.

Clarkin, J. J. Effects of the operational environment on performance measurement. Proceedings, Conference on Productivity Enhancement: Personnel Performance Assessment in Navy Systems, Navy Personnel Research and Development Center, San Diego, CA, October 1977.

Dalkey, N. and Helmer, F. An experimental application of the DELPHI method to the use of experts. Mgmt. Sci., 1963, 9, 458-467.

Fineberg, M. L. Navigation and flight proficiency under NOE conditions as a function of aviator training and experience. Proceedings, Human Factors Society, 18th Annual Meeting, 1974, 249-254.

Glaser, R. et al. A study of some dimensions of team performance. Tech. Rep. Contract N7onr-37008, American Institute for Research, Pittsburgh, PA, September 1955.

Meister, D. et al. Training effectiveness evaluation of naval training devices. Part II: A study of device 2F66A (S-2E Trainer) effectiveness. NAVTRADEVCEN 69-C-0322-2, Naval Training Device Center, Orlando, FL, July 1971.

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

Meister, D. Where is the system in the man-machine system? Proceedings, Human Factors Society, 18th Annual Meeting, 1974, 287-292.

Meister, D. Heresies: brief essays on Human Factors. Unpublished report, Navy Personnel Research and Development Center, San Diego, CA, March 1977.

Meister, D. Behavioral Foundations of System Development. New York: Wiley, 1976.

Parsons, H. M. Man-Machine System Experiments. Baltimore: Johns Hopkins Press, 1972.

Peters, G. A. and Hall, F. S. Missile system safety. Report ROM 3181-1001, Rocketdyne Corporation, Canoga Park, CA, March 1963.

Smode, A. F. et al. The measurement of advanced flight vehicle crew proficiency in synthetic ground environments. Report MRL-TDR-62-2, Behavioral Science Lab., Aerospace Medical Division, Wright-Patterson AFB, Ohio, February 1962.

ABOUT THE AUTHOR

Dr. Meister received his Ph.D. in psychology in 1951 from the University of Kansas. He has spent 8 years working at various times for both the Army and the Navy, and 16 years in industry both in research and application positions with General-Dynamics Astronautics and the Bunker-Ramo Corporation. He is the author of 3 books on human factors and was President of the Human Factors Society in 1974-5. He has written many papers in the field of human performance reliability and lectures occasionally on the topic.

PERFORMANCE TESTING IN INSTRUCTIONAL SYSTEMS

John Brock
Navy Personnel Research and Development Center
San Diego, California

ABSTRACT

The role of performance testing in the design and evaluation of instructional systems is discussed. The Instructional Systems Design (ISD) process is reviewed in detail. A model for instructional system evaluation using performance measures is suggested and specific R&D proposals are briefly discussed.

INTRODUCTION

Background

All three services are embarked on the systematic design of instructional systems (e.g., Haverland, 1976; Scanland, 1974; Ricketson, Wright and Schultz, 1971). Attempts to develop an instructional design methodology which would be acceptable to Army, Navy, and Air Force instructional designs have consistently failed (e.g., see Montemerlo and Tennyson, 1976, for a more nearly complete discussion of this phenomenon). However, it is generally agreed that job relevant training which exploits modern instructional technology is a desirable goal (Chief of Naval Education and Training, 1975; Brock, 1977a).

Called variously course design procedure, systems approach to training, or instructional systems design (ISD), the essential systems approach takes the training course developer from a set of job tasks, through the development of behavioral objectives, to the conduct of a job relevant training program (Brock and DeLong, 1975; Montemerlo and Tennyson, 1976; Freitag and Mitzel, 1977).

Purpose

This paper will review the state of the art in instructional systems design (ISD) technology and the significant role of performance testing in the design process. The use of performance tests to evaluate extant instructional systems will be discussed in some detail. The paper will conclude with the writer's suggestions for future R&D programs directed at improving the instructional process through the use of performance measures.

INSTRUCTIONAL SYSTEMS DESIGN

Assumptions

Instructional systems design (ISD) in the military is based on several assumptions which are seldom articulated but omnipresent. The overriding assumption is one of policy: it is in the best interests of the military to train men and women to perform specific jobs or clusters of tasks which must be performed in the work environment.

PRECEDING PAGE BLANK-NOT FILMED

A corollary of this assumption is that it is the job which is to be performed immediately after the schoolhouse instruction for which training must be given. Although this policy is open to question (e.g., Brock, 1977b), this paper will treat it as a given.

A third assumption is that any person of normal intelligence, with a modicum of training, can design instruction. This assumption grows from an even more deeply rooted one: nearly anybody can teach.

Yet another set of assumptions is that (1) there is a single way to design training and (2) it is capable of being proceduralized. These assumptions are not supported by any evidence (Montemerlo and Tennyson, 1976).

Finally, there is an assumption that the terminology used to describe ISD operations conveys essentially the same information to all who come in contact with it. Once again most of the evidence supports a contrary point of view: the ISD terminology has an almost person-specific meaning. For example, the term "task analysis" can mean anything from a computer analysis of a job category (who does it, how many, how often) to an in-depth logical analysis of a particular job.

A major point of this paper will be that many of the current problems with performance testing in instructional systems stem from the above faulty assumptions, rather than from an insufficient technology.

The military services must rely on their overall manpower pools for their instructor cadre. However, better selection of instructors must be instituted, instructor training must be greatly improved, and adequate job aiding must become available. There is no reason to believe that an enlisted man or officer who is expert in a particular job area will have the necessary skills to either design or conduct training on those jobs. To further expect them to develop job performance is ludicrous.

This leads to the fourth set of assumptions: that some sort of fully proceduralized job aid can be developed so that the method of instructional design can be performed by a typical enlisted or officer military instructor. As Montemerlo and Tennyson (1976) point out, there is neither empirical nor theoretical support for this assumption. This, of course, has not inhibited anyone from developing fully proceduralized instructional design procedures (e.g., Rundquist, 1970; USAF AFP 50-58, 1974; NAVEDTRA 106A, 1975).

As discussed above, there is the assumption underlying current ISD techniques that the technology has an established vocabulary. Since there is significant evidence that this is not the case, a limited number of ISD terms will be defined. The terms will be limited to those which apply directly to performance testing design and conduct.

There are three terms which are used as if they have meaning but which are used interchangeably so often as to obsure rather than clarify sense. They are "front-end analysis," "task analysis," and "training analysis." For the purposes of this paper, "training analysis" will not be used. "Front-end analysis" is used to describe the entire analytic process that goes into producing some end product; in this paper, it is the analytic techniques which result in performance tests within an instructional systems context.

A component, or phase, of a front-end analysis is a task analysis. As used in the present discussion, this refers to an analytical dissection of specific jobs into hierarchial arrangements. Typically, such analyses begin with global task statements (e.g., "locates a malfunction in the electrical system of an automobile") and reduce down to unitary behaviors (e.g., "removes distributor cap"). A complete description of this technique, with examples, can be found in Rundquist (1970), Brock and DeLong (1975), and Brock (1977a). It will be further discussed below as it applies to the development of performance tests. Other terminology will be defined as it occurs in the paper.

Performance Testing in ISD

It is neither an overstatement nor an oversimplification to state that performance tests lie at the heart of the instructional system design process. Figure 1 is taken from the Chief of Naval Education and Training manual on ISD. The figure is a block diagram of the major steps in ISD. Note that the third block is, "Construct job performance measures." In other words, the instructional system flows from the performance tests which, as can also be seen in Figure 1, flow in turn from what is called in this paper the task analysis.

The point appears obvious. If the Services are going to have performance oriented training, then the goal of instruction is to modify performance. In order to find if that performance has been appropriately modified, that performance must be measured. In other words, performance testing is the beginning and the end of the instructional process.

The Behavioral Objective

One of the few agreed upon definitions in instructional technology is that of the behavioral objective: "(It) is an intent communicated by a statement describing a proposed change in a learner--a statement of what the learner is to be like when he has successfully completed a learning experience" (Mager, 1962, p. 3). Every behavioral objective must either explicitly or implicitly delineate a specific behavior, the conditions under which the behavior is to be performed, and the standard to which the behavior is to be performed. In other words, a behavioral objective is the description of a performance test.

Visualize an instructional process as a road with a beginning (training prerequisites) and an end (terminal behavior of the students). The milestones along the road are the performance objectives of the course; the student progresses on the road by passing each objective.

Task Analysis

How one derives performance objectives is, of course, the critical question in ISD. The writer's bias is clear: there is currently no prescriptive methodology sufficiently detailed that a typical military instructor can conduct a thorough front-end analysis. Therefore, the following discussion assumes an instructional design team made up of (1) experts on the tasks to be trained and (2) instructional technologists.

Assumptions of any task analysis are that job behavior is organized and that clever analysis of a job will uncover that organization. Recently, the trend has been to break down jobs into increasingly more specific tasks. These breakdowns are

typically referred to as learning or instructional hierarchies (e.g., Brock, 1977a; Malone, DeLong, Farris and Krumm, 1976; Brock and DeLong, 1975).

For the designer of performance tests, a well defined hierarchy of tasks can aid him in determining precisely what inferences can be drawn from his tests. For instance, if a task at the top of a hierarchy is to be tested, one can reasonably infer that all the tasks necessary to support that task have been learned, e.g., if a technician locates a series of malfunctions in an automobile's electrical system, one can infer that he or she can remove a distributor cap. However, the inverse is not true: it cannot be inferred that a technician can locate malfunctions in an electrical system because he or she can remove a distributor cap, change spark plugs, or perform other equally low level tasks.

The same is true regarding knowledge and skills. If a person can troubleshoot this electrical system, it seems reasonable to infer that the person has sufficient knowledge (e.g., electrical theory, general electrical system principles) for the job. However, just because the technician can answer a series of questions about electrical theory does not warrant the inference that he or she is a competent troubleshooter. This point has been discussed in earlier papers in this symposium (e.g., Crawford and Brock, 1977). It is the failure to understand and discern these hierarchical relationships within task clusters which has created much bad instruction and many irrelevant performance measures.

The overt skill hierarchy is reasonably easy to derive. Many jobs are either sequenced (e.g., missile firing) or repetitive (e.g., assembling electronic components). Higher level skills (e.g., flying an airplane) are more difficult to organize but the process is essentially the same. The writer has had good experience by simply asking the question of the job expert, "What does he have to do to accomplish X?" This question is asked until a reasonably low level of behavior is reached. "Reasonable" is extraordinarily difficult to define, but agreement as to what is reasonable in a specific setting is usually quite quickly reached (Brock and DeLong, 1975; Brock, 1977a).

The difficulty in hierarchical derivation comes when the nonobservable skill and knowledge elements must be articulated. These become inportant to the instructional designer not only because he has to identify what he wants to teach, but also because it is the primary guide to what and how to test. For instance, to use what is fast becoming a hackneyed example, let us assume that one wishes to determine if a technician can locate a series of malfunctions in an automobile electrical system. Let us further assume that we do not have an automobile, or at least one into which we can insert known malfunctions. The earlier discussion warned what cannot be inferred; however, by a clear analysis of the skill and knowledge element hierarchy supporting troubleshooting of the electrical system, tests can be designed which can measure up the hierarchy to some specific level. By this technique, the trainer at least has precise understanding of what the student can do.

Figure 2 is an example of a task hierarchy for a pilot in the F-14 fighter aircraft. Measurement of performance in a MACH 1.5 jet aircraft is difficult at best; earlier papers in this conference make that clear (Vreuls, 1977). However, if a test designer could test the second level of tasks in the hierarchy in Figure 2 (PSUS 1-1-1 through PSUS 1-6-1), could not a reasonable inference be made about the pilot's ability to establish an initial search configuration?

Testing performs many functions in an instructional program; the evaluation of instructional system function will be discussed in detail below. However, if one looks at Figure 2, one can perceive a series of performance tests which serve as milestones and diagnostic instruments for the students, location markers for the course managers, and trouble lights for the instructional designer. Additionally, it is clearly preferable to identify that a student cannot set his radar in the proper mode prior to his flying the airplane.

Several theoreticians have attempted to build hierarchical learning categories (e.g., Bloom, 1956; Krathwohl, Bloom and Masia, 1964; Gagne, 1970; Merrill, 1971a; Markle and Tieman, 1973). If a task can be fitted to a category, then it can be fitted into a prestructured hierarchy as well. If the characteristics of a particular category are well defined, then a technology on how to (1) train the behavior and (2) measure the behavior can be evolved.

The writer has devised a scheme for translating observable task behaviors into one of several learning category systems (Brock, 1976, 1977a). Figure 3 lists the behavioral categories as they were applied in the F-14 ISD process; Figure 4 is an algorithm which was used to assign tasks to a particular category.

The task category system can also be used to assign specific tasks to particular testing techniques. A high-fidelity simulator is not necessarily the best place --or even an adequate place--to measure some kinds of task performance. Behavioral categories 6 and 7 appear to be best tested in a classroom or learning carrel which minimize irrelevant cues. For example, solving an intercept geometry problem is a highly cognitive skill that requires much practice. To measure this skill while "flying" a simulator (with all its attendant problems) is unreasonable. This type of classroom cognitive test process has been successful with surface Navy officers (McCutcheon and Brock, 1971; Brock, 1972).

The tasks in categories 4 and 5 are those that require the high fidelity of a Weapon System Trainer for testing (e.g., flight maneuvers for the pilot and complex search procedures for the Navy Flight Officer).

Categories 1, 2, and 3 will be best tested in dynamic part-task trainers that only need to simulate key subsystem cues and responses. The kind of part-task training developed by Crawford (1976) for the S-3A Viking aircraft allows efficient testing of the skills in these categories.

The point of all this activity is to bring to the test design stage a set of job tasks and their supporting skill and knowledge elements; in other words, the front-end analysis gives the instructional design team the data base from which they can proceed.

ISD Test Design

Figure 5 presents the CNET ISD flow chart for the construction of job performance measures. The ten steps which are shown stem from a task analysis such as the one described above. The CNET guide states, "Once the decision has been made as to which tasks will be trained, it is necessary to construct performance measures to test whether individuals can perform the tasks. These job performance measures (JPMs) become the fundamental basis for the development and control of training since they are the measure of the success of training" (NAVEDTRA 106A, 1975, p. 156).

The CNET model makes an important distinction between job performance measures (JPMs) and job performance tests. A JPM is written at the task level and can measure one or more complete tasks. A job performance test is a test used to determine whether or how well an individual can perform a job. It is the writer's belief that too much is made of this distinction in an instructional system which is presumed to be job relevant. In such a system, the job performance tests should simply be aggregates of JPMs.

Many of the issues raised in this conference are discussed in the CNET model: predictive validity, physical fidelity, simulation, and unitary versus multiple tasks. Although the discussion in the model may be adequate for the intended readership, it falls short of being prescriptive.

It does not seem appropriate to detail each step in the CNET model of JPM design. Suffice to say that the ten steps (Figure 5) represent a reasonable approach which is discussed in detail in the model. Even with examples, however, the criteria for making decisions about what and how to test are frustratingly vague. For instance, constraints such as time, money, and manpower are discussed as reasons for a JPM to be less than actual job performance. However, how to maximize one's test within these constraints is not discussed. The section on measuring product or process is only slightly more illuminating.

The writer will discuss the R&D implications of this discussion in a later section of this paper. For the moment, it will suffice to say that if the hierarchical front-end analysis described above is performed, many of the traditional barriers to deciding what to measure will be overcome. Deciding how to measure those behaviors will also be facilitated, but not to the extent that the "what" question is answered.

EVALUATION OF INSTRUCTIONAL SYSTEMS

The evaluation of instructional systems has been the subject of books (e.g., Cronlund, 1968), complete issues of journals (e.g., Tiemann, 1976), and running series in journals (e.g., Elsbree and Howe, 1977). To review all the literature overwhelms the writer, it not the reader. Therefore, this section of the paper will touch on some of the models for the evaluation of instruction and discuss performance testing in the context of those models.

On the surface, one is tempted to ask, "So, what's the problem?" The Services want to train performance; therefore, the quality of a particular instructional system is reflected in how well the students perform the objectives of the system.

As this conference has pointed out several times, the problem is how to measure performance. It has been suggested above that the failure to adequately specify what to measure has also created problems. If the how and the what are resolved, is the problem of evaluating instructional systems resolved? Is further discussion necessary?

As it turns out--and this should come as no surprise--some further discussion is appropriate. A model for evaluation is still lacking; primarily, there is a lack of prescription for where evaluation of the system fits, when evaluation should be done, and what should be done with the results of any evaluation.

Brethower and Rummler (1977) propose three general systems models which will be discussed in the context of both performance testing and the military system. Figure 6 presents these three systems.

The ballistic system is of little concern to the designer of performance measures, since there is no need to measure the system output. The guided system fits the more traditional military training system. Assume that the circle with the X in it is the measurement point for system evaluation. In performance oriented instruction, this would be the place for the evaluative performance test. Number of students meeting some predetermined percentage of objectives has been the traditional quantitative evaluation of this system. In the Navy, an instructional system has traditionally been judged adequate if 90 percent of the students meet 90 percent of the objective. The danger of this kind of evaluation will be discussed below.

The adaptive system provides for two points of evaluation; one immediately after training, a second on the job. It is this last model that most fully exploits the front-end analysis technology discussed above. The first evaluation point measures how well the system objectives are being met; the second point measures how well the instructional system is meeting the needs of the operational system. It is only in an adaptive system, with its appropriate feedback loops, that the instructional system has the necessary information to change its objectives.

The danger in both the guided and adaptive systems is the seductiveness of quantitative data such as the 90/90 criterion cited above. Let us assume that in a pilot training program there are one hundred objectives. Let us further assume that 95 percent of the pilot students are meeting better than 90 percent of the course objectives. With such a gross measure, no trouble in the system is indicated. However, a careful look at what objectives are not being met could identify symptoms of an ailing instructional program, e.g., 90 percent of the students not being able to eject.

Performance testing of students at the completion of training only makes sense if a qualitative analysis of the performance tests is made. For a complete discussion of training effectiveness evaluation, the reader is referred to Semple (1974). He refers to four levels of instructional system evaluation, which are based upon work done by Jeantheau (1971), as summarized by Blaiwes, Puig and Regan (1973). The first three of these levels can be performed within the system; the guided and adaptive systems are equally able to be evaluated at these levels.

The first level of evaluation is qualitative. Content, methods, media, and procedures are examined in terms of particular objectives being met or not met. Sources other than performance measures are used; however, the writer views them as secondary.

The second level of evaluation is non-comparative performance measurement. Essentially, this means testing the student's performance before, during, and after training. The degree of improved student performance is, presumably, highly correlated with the quality of the instructor.

The third level involves comparative measurement. Two instructional systems with the same objectives are compared on how well the objectives are met. Typically, this kind of comparison would only be possible for small units of instruction or alternative training devices.

The fourth level is only available with the adaptive model: transfer of training --the comparative measurement of task performance in an operational situation. Note that on the face of it, this is a different definition of transfer than one normally encounters.

Typically, transfer is an attempt to measure the effect of learning one task on learning a different task (Roscoe, 1971). A typical design of such a study might have one group of subjects learn a pursuit rotor task before, say, a finger maze task; a second group would learn only the finger maze. If the pursuit rotor group learns the finger maze more quickly or better on some dimension, transfer is implied. Semple (1974) does not make clear that this is his measuring of transfer. However, the implication is that job environments are different enough from training environments that transfer refers to the degree the graduate can perform on-the-job because of his learned performance in the instructional system.

What is herein proposed is the adaptive model of instructional system evaluation based upon at least two measures of performance: (1) at the completion of the instruction and (2) in the job environment. The first measure is an indicator of the internal state of the system--is the system meeting its objectives? The second measure is, first of all, an indicator of how well the system is meeting the needs of the consumer and, secondly, how much of the instruction is staying with the graduate of the system. Without both measures, information about instructional systems will continue to lack explicitness and, therefore, to provide the feedback necessary for appropriate corrective actions to be taken.

Most of the literature in performance measurement in instructional systems addresses evaluating student performance in order to discover something about the student (e.g., Merrill, 1971; Glaser and Klaus, 1962). This discussion has attempted to point out issues in using these same tests to evaluate instruction. There are several areas left undiscussed; experimental designs come to mind and, as a subset of that, statistical techniques which could apply to performance evaluations of instructional systems.

For the former, the reader is referred to Blumenfeld and Holland (1971) who make an ardent appeal for control groups and Brenthower and Rummler (1977) who offer alternatives to the control group design. For the latter, the writer suggests any good statistics text (e.g., Edwards, 1968) or, for a Bayesian procedure, Hambleton and Novick (1973).

The writer's own experience is that one does as well as he can. Instructional systems and devices get evaluated, often by measuring the performance of students and nonstudents. To cite a few, performance measurement techniques have been used to evaluate electronic maintenance training systems (e.g., Daniels, Datta, Gardner and Modrick, 1975; Wright and Campbell, 1975), driving simulators (Bishop, 1967; Edwards, Hahn and Fleishman, 1969); a welding simulator (Abrams, Safarjan and Wells, 1973); and aircrew training devices (Cream, Eggemeir and Klein, 1975). While none of these studies meets the elegance of formal experiments, their findings appear valid and much is known of the systems and devices of concern.

The need still exists for a systematic approach to instructional system design. Typically, system evaluation designs are driven by constraints rather than by needs. As the need for more efficient and effective instructional systems is felt, improved evaluation models exploiting performance measurement techniques should be forthcoming.

CONCLUSIONS AND RECOMMENDATIONS

Much is known about designing instructional systems; the emphasis on designing a system from valid performance tests is healthy. A prescription for the design of performance tests is lacking and research in this area is urgently needed.

As a follow-up to the above, a major R&D effort should be instigated on the development of job aids--possibly computer supported--for the designers of performance tests.

The requirement for two performance test points to evaluate instruction is manifest. Anderson, Laabs, Pickering and Winchell (1977) have proposed a comprehensive job proficiency assessment system. This proposal deserves the highest attention and should be supported.

And finally, performance test design based on front-end analysis and instructional system evaluation must be treated by ISD technicians as the beginning and end of a single process. Integration of instructional design functions will not only eliminate duplication of effort but will produce congruent instruction, measurement, and jobs.

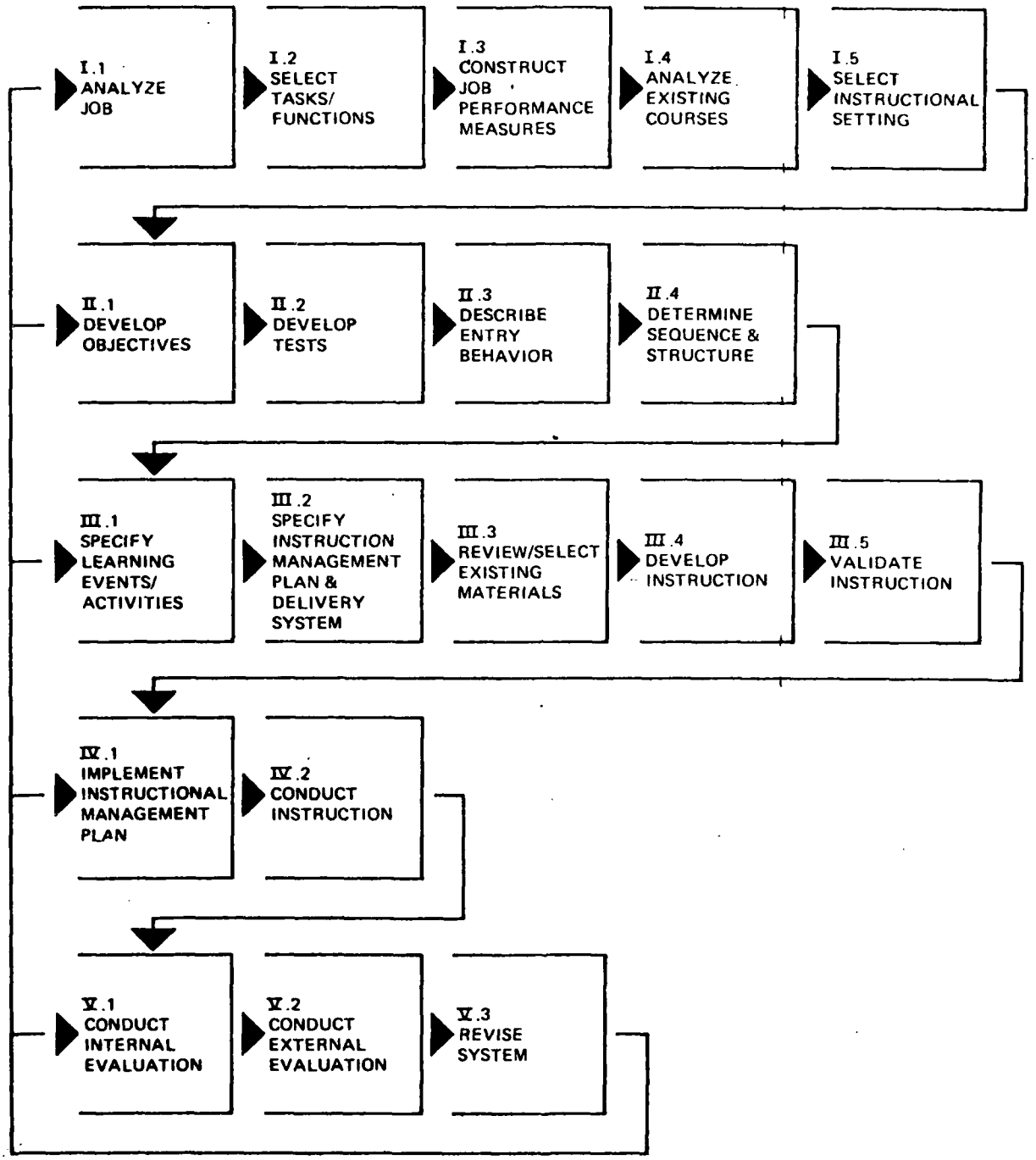


Figure 1. The CNET ISD Process (NAVEDTRA 106A, 1975)

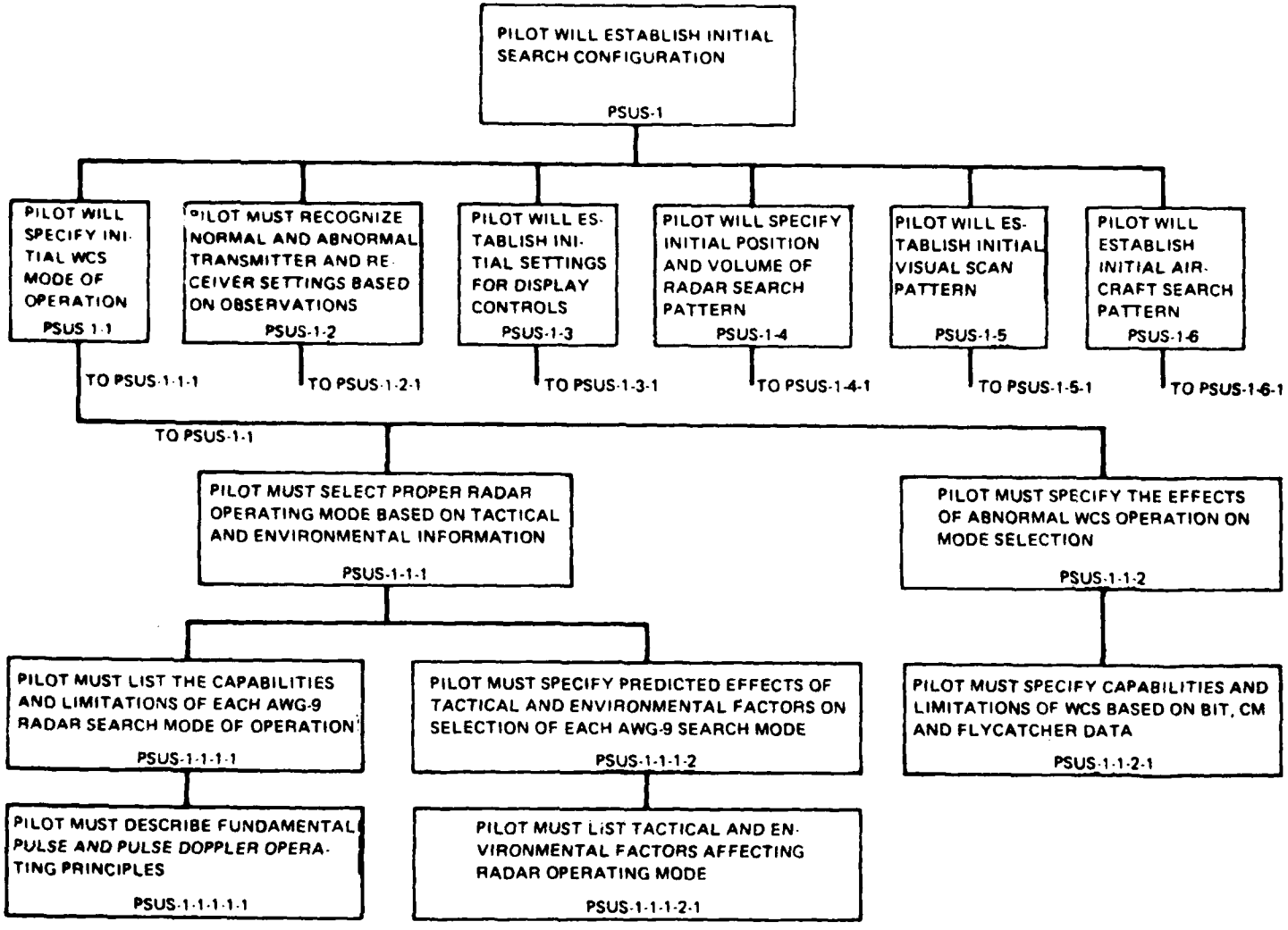


Figure 2. Example of F-14 pilot task analysis hierarchy (Brock, 1977a).

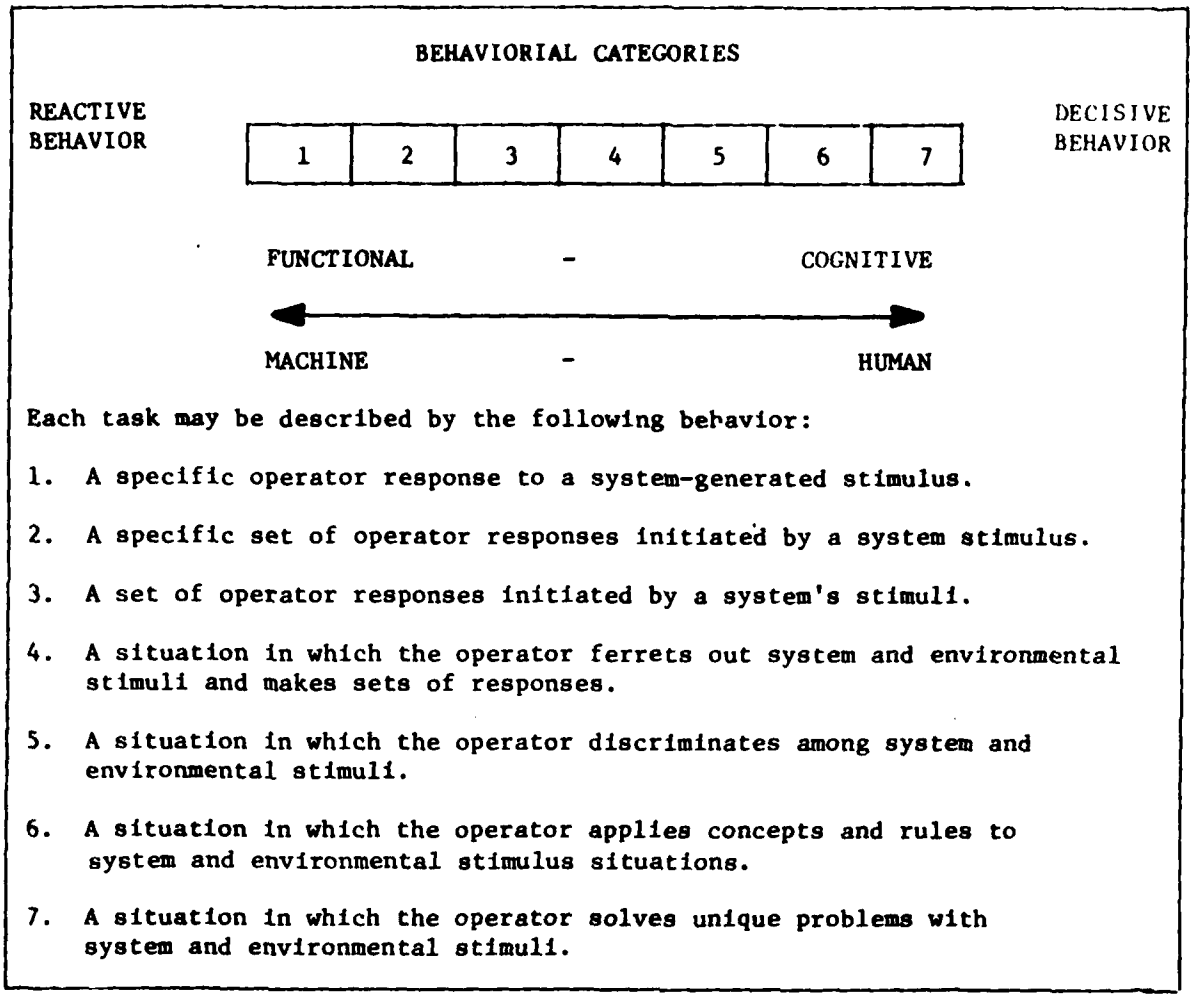


Figure 3. Behavioral category for F-14 weapons system operation (Brock, 1976).

BEHAVIORAL CATEGORY	ASK THE QUESTION:
	Does the system produce a simple stimulus and does the operator make a single response?
If the answer is:	
YES, then use 1	NO, then ASK THE QUESTION:
	Does the system produce a simple stimulus and does the operator make a sequence of responses?
If the answer is:	
YES, then use 2	NO, then ASK THE QUESTION:
	Does the system produce a simple stimulus followed by an operator response (or responses) which produces a second stimulus which produces a second response, etc.?
If the answer is:	
YES, then use 3	NO, then ASK THE QUESTION:
	Does the operator actively interact with the system and environment, attending to selected stimuli and making appropriate responses?
If the answer is:	
YES, then use 4	NO, then ASK THE QUESTION:
	Is the operator presented with discrete sets of system and environmental stimuli and must he distinguish among them?
If the answer is:	
YES, then use 5	NO, then ASK THE QUESTION:
	Can the operator completely respond to the environmental and system stimuli by applying set rules and concepts?
If the answer is:	
YES, then use 6	NO, then ASK THE QUESTION:
	Must the operator make unique responses to an infinite combination of environmental and system stimuli?
If the answer is:	
YES, then use 7	NO, then REEVALUATE THE BEHAVIOR.

Figure 4. Behavioral category selection algorithm (Brock, 1977a).

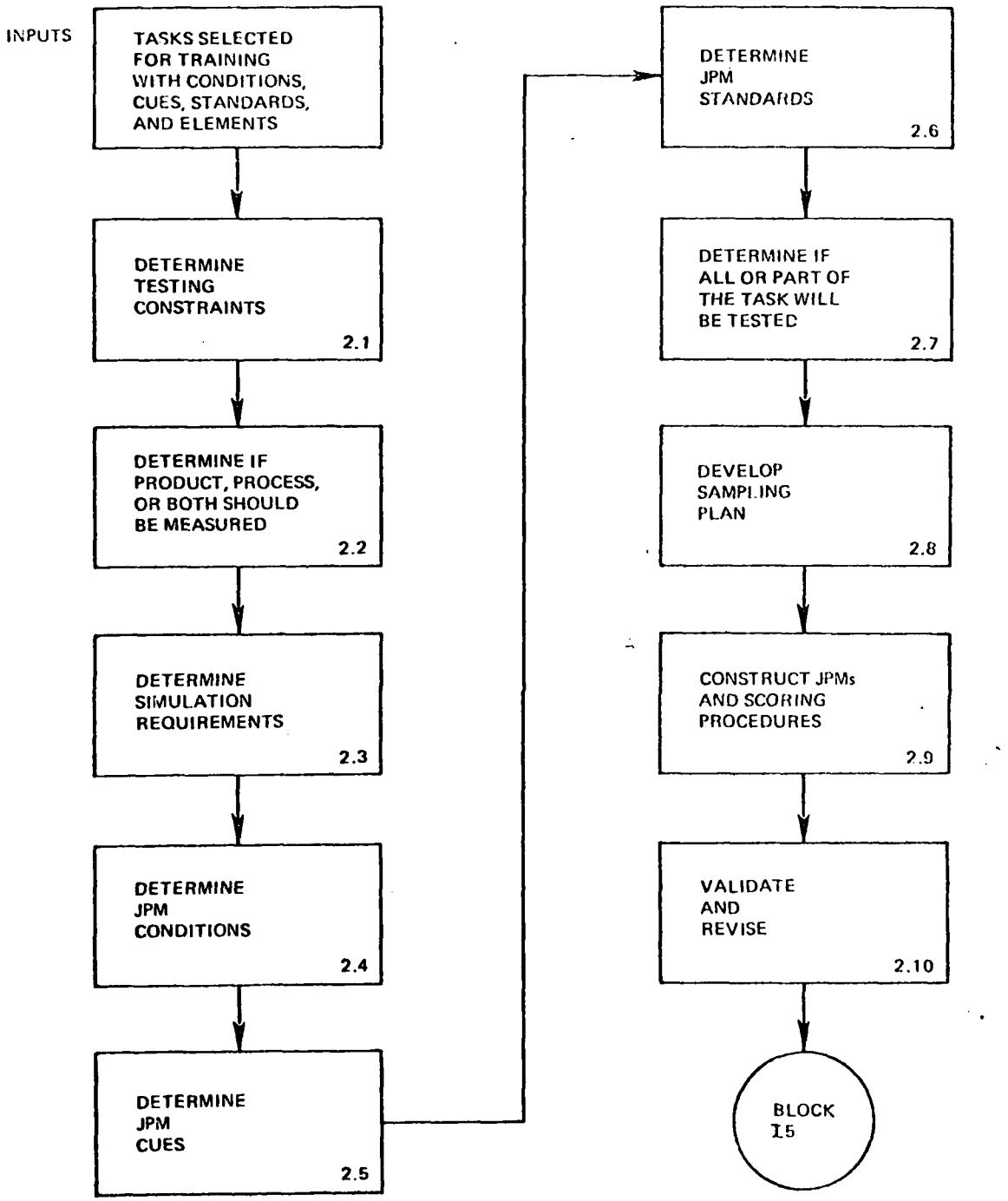


Figure 5. Flowchart of CNET ISD Block I.3: Construct Job Performance Measures (NAVEDTRA 106A, 1975, Figure I.18).

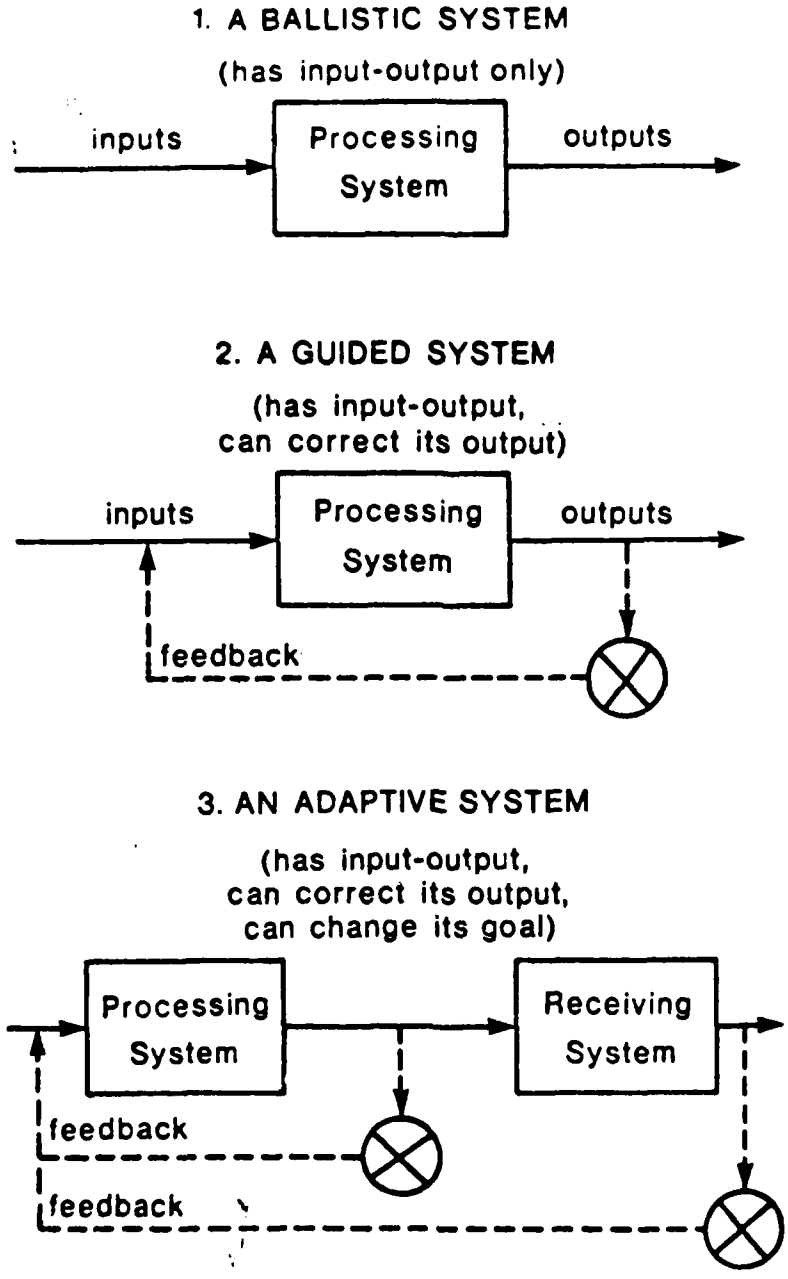


Figure 6. General Systems Model (Brethower and Rummler, 1977, p. 106).

REFERENCES

Abrams, M. L., Safarjan, W. R. and Wells, R. G. Description and preliminary training evaluation of an arc welding simulator (NPTRL SRR 73-23). San Diego: Navy Personnel and Training Research Laboratory, June 1973.

Anderson, A. U., Laabs, G. J., Pickering, E. J. and Winchell, J. D. A personnel readiness training program: Final report (NPRDC TR 77-39). San Diego: Navy Personnel Research and Development Center, August 1977.

Bishop, R. W. Evaluating simulator instruction for accomplishing driver education objectives. Research Review, 1967, 11, 12-17.

Blaiwes, A. S., Puig, J. A. and Regan, J. J. Transfer of training and the measurement of training effectiveness. Human Factors, 1973, 15, 525-535.

Bloom, B. S. (Ed.). A taxonomy of educational objectives: Handbook I, the cognitive domain. New York: Cougmans, Green and Co., 1956.

Blumenfeld, W. S. and Holland, M. G. A model for the empirical evaluation of training effectiveness. Personnel Journal, August 1971, pp. 637-640.

Brethower, K. S. and Rummler, G. A. Evaluating training. Improving Human Performance Quarterly, 1977, 5, 3-4; 103-120.

Brock, J. F. Development of two models for improvement of a combat information center watch officer course: A proposal for implementation (Research Memorandum SRM 73-1). San Diego: Naval Personnel and Training Research Laboratory, 1972.

Brock, J. F. Development of a task category system for the design of air crew training. Paper presented at the Fifth Psychology in the Air Force Symposium, United States Air Force Academy, Colorado, April 1976.

Brock, J. F. Instructional decision making in the design of operator training: An eclectic model (NPRDC TR 77-31). San Diego: Navy Personnel Research and Development Center, May 1977a.

Brock, J. F. Simulation in maintenance training: Now that I've thrown out the bath water, where is the dear baby? Paper presented at the 1977 American Educational Research Association, New York, New York, April 1977b.

Brock, J. F. and DeLong, J. L. Design and conduct of a mechanical maintenance training program with an annual flow rate of 11,000 trainees: A review. In W. J. King and J. S. Duvo (Eds.), New Concepts in Maintenance Trainers and Performance Aids (NAVTRAEQUIPCEN IH-255). Orlando, FL: Human Factors Laboratory, Naval Training Equipment Center, October 1975, pp. 103-116.

Chief of Naval Education and Training. A plan for centralized management of instructional systems development within the Naval Education and Training Command, August 15, 1975.

Chief of Naval Education and Training. Interservice procedures for instructional systems development (NAVEDTRA 106A). Pensacola, FL: Chief of Naval Education and Training, 1975.

Crawford, A. M. Low-cost training using interactive computer graphics. Paper presented at the Fifth Psychology in the Air Force Symposium, United States Air Force Academy, Colorado, April 1976.

Crawford, A. M. and Brock, J. F. Using simulators for performance measurement. Paper presented at the Symposium on Productivity Enhancement: Personnel Performance Assessment in Naval Systems, San Diego, October 1977.

Cream, B. W., Eggemeier, F. T. and Klein, G. A. Behavioral data in the design of aircrew training devices. Proceedings of the Human Factors Society 19th Annual Meeting, Dallas, TX: October 1975.

Daniels, R. W., Datta, J. R., Gardner, J. A. and Modrick, J. A. Feasibility of automatic electronic maintenance training (AEMT) Volume I--Design development and evaluation of AEMT/ALQ-100 demonstration facility (NADC 75176-40). Warminster, PA: Naval Air Development Center, May 1975.

Department of the Air Force. Handbook for designers of instructional systems (AFP 50-58). Washington, DC: Headquarters, United States Air Force, 1974.

Edwards, A. L. Experimental design in psychological research. New York: Holt, Rinehart and Winston, Inc., 1968.

Edwards, D. S., Hahn, C. P. and Fleishman, E. A. Evaluation of laboratory methods for the study of driver behavior: The relation between simulator and street performance. American Institutes for Research: Washington, DC, R69-7, May 1969.

Elsbree, A. R. and Howe, C. An evaluation of training in three acts. Training and Development Journal, August 1977, pp. 12-19.

Freitag, M. and Mitzel, W. Navy instructional program development process evaluation plan. Paper presented at the 1977 American Educational Research Association, New York, New York, April 1977.

Gagne, R. M. The conditions of learning (2nd Ed.). New York: Holt, Rinehart and Winston, 1976.

Glaser, R. G. and Klaus, D. J. Proficiency measurement: Assessing human performance. In Robert M. Gagne (Ed.), Psychological Principles in System Development. New York: Holt, Rinehart and Winston, 1962, 419-474.

Gronlund, N. E. (Ed.). Readings in measurement and education. New York: The Macmillan Company, 1968.

Hambleton, R. E. and Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, Fall 1973, 10(3), 159-170.

Haverland, E. M. Transfer and use of training technology in Air Force technical training: A model to guide training development (HumRRO FR-ED 76-43). Alexandria, VA: Human Resources Research Organization, 1976.

Jeantheau, C. G. Handbook for training systems evaluation (NAVTRADEVCEN 66-C-0113-2). Orlando, FL: Naval Training Device Center, January 1971.

Krathwohl, D. R., Bloom, B. S. and Masia, B. A taxonomy of educational objectives: Handbook II, the affective domain. New York: David MacKay and Company, Inc., 1964.

Mager, R. F. Preparing instructional objectives. Palo Alto: Fearon Publishers, Inc., 1962.

Malone, T. B., DeLong, J. L., Farris, R. and Krumm, R. L. Advanced concepts of Naval Engineering Maintenance Training (NAVTRAEQUIPCEN N-61339-74-C-0151). Orlando, FL: Naval Training Equipment Center, May 1976.

Markle, S. M. and Tiemann, P. W. The learning process. Chicago: Tiemann Association, 1973.

McCutcheon, R. E. and Brock, J. F. Effects of establishing a conceptualization context for learning monitoring and evaluating tasks (NPTRL Res. Rept. 72-4). San Diego: Naval Personnel and Training Research Laboratory, 1971.

Merrill, M. D. Necessary psychological conditions for defining instructional outcomes. Educational Technology, August 1971a, pp. 34-39.

Merrill, M. D. (Ed.). Instructional Design: Readings. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1971b.

Montemerlo, M. D. and Tennyson, M. E. Instructional systems development: Conceptual analysis and comprehensive bibliography (NTEC IH-257). Orlando, FL: Naval Training Equipment Center, February 1976.

Ricketson, D. S., Schultz, R. E. and Wright, R. H. Review of the CONARC systems engineering of training program and its implementation at the United States Army Aviation School. Fort Rucker, AL: Human Resources Research Organization, April 1970.

Roscoe, S. N. Incremental transfer effect. Human Factors, 1971, 13(6), 561-567.

Rundquist, E. A. Job training course design and improvement (2nd ed.)(NPTRL RR 71-4). San Diego: Naval Personnel and Training Research Laboratory, 1970.

Scanland, W. Developing the instruction. Campus, April 1974, pp. 14-16.

Semple, C. A. Guidelines for implementing training effectiveness evaluations (NAVTRAEQUIPCEN 72-C-02-9-3). Orlando, FL: Naval Training Equipment Center, April 1974.

Tiemann, P. W. An evaluation issue. Improving Human Performance Quarterly, Fall-Winter 1976, 5, 3-4, introduction to a special issue.

Vreuls, D. Performance measurement of aircrew personnel. Paper presented at the Symposium on Productivity Enhancement: Personnel Performance Assessment in Naval Systems, San Diego, October 1977.

Wright, J. and Campbell, J. Evaluation of the EC-II programmable maintenance simulator in T-2C organizational maintenance training (NADC 75083-40). Warminster, PA: Naval Air Development Center, May 1975.

ABOUT THE AUTHOR

John F. Brock has been an Education Specialist at the Navy Personnel Research and Development Center and its predecessors for 10 years. Prior to that, Mr. Brock served as a Navy officer on board USS PERKINS (DD-887), USS WRIGHT (CC-2) and the Fleet Anti-Air Warfare Training Center, San Diego. He has conducted research into instructional systems design, programmed instruction, shipboard training systems, engineering maintenance training, aircrew training development and most recently, has begun a R&D program on developing a Life Cycle Costing Model for instructional systems. Mr. Brock has done graduate work in experimental psychology at San Diego State University. He is a member of the Human Factors Society, The American Education Research Association, The National Society for Performance and Instruction, and the Society for Applied Learning Technology.

ON THE MEASUREMENT OF MAN-MACHINE PERFORMANCE

Robert R. Mackie
Human Factors Research, Incorporated
Goleta, California

ABSTRACT

Some fundamental characteristics of man-machine systems that affect performance measurement methodology are considered. Each of the following elements of man-machine performance measurement is then discussed:

1. Deciding what one wishes to learn from the man-machine performance measurement in terms of broad system objectives.

2. Identifying appropriate operational criteria.

3. Designing a comprehensive test scenario.

4. Selecting the most appropriate test environment.

5. Selecting and training the test personnel.

6. Developing a suitable data collection methodology.

7. Analyzing and synthesizing the data into operationally meaningful evaluations.

SOME FUNDAMENTAL CHARACTERISTICS OF MAN-MACHINE SYSTEMS THAT AFFECT PERFORMANCE MEASUREMENT

Man-Machine Systems Are Goal-Oriented

There are a number of characteristics of man-machine systems that influence the approach to performance measurement in these systems. The most fundamental is that man-machine systems are by nature goal-oriented. Meister (1976) has pointed out that, "The common element in all definitions of a man-machine system is the concept of purposiveness. Since the man-machine system is an artificial creation, its characteristics depend on the purpose of its creator." It follows that, to appropriately measure man-machine systems performance, one must identify criteria that reflect the goals of the system. A problem, however, is that there are many different goals associated with different levels of the system. The most general or highest level system objectives often do not directly reflect very much about the performance of the subsystems at lower levels within the system.

Meister has also pointed out that, because the operator is a subsystem of man-machine systems, the overall goals of the system must control his behavior. He is effective when he implements these goals and ineffective when he does not. The significance of his behavior in terms of its effect on the overall system can be determined only in relation to these goals.

PRECEDING PAGE BLANK-NOT FILMED

Man-Machine Systems Are Hierarchical

A second feature that influences performance measurement in man-machine systems is that, except for the unique case of a single operator and machine operating in isolation, man-machine systems contain a hierarchy of subsystems, each higher level being composed of systems at lower levels. As Meister has noted, the concept of system level is important because higher-order subsystems may have properties of complexity and dependency that are not found in lower-level subsystems. Further, the interaction between different systems levels is obviously of considerable consequence for overall system functioning.

Man-Machine Systems Vary in Determinacy

A third aspect of man-machine systems that significantly impacts performance measurement methodology is the degree of determinacy associated with the system. To quote Meister (1976) again:

> The theme that runs through various systems is one of probability; in determinate systems the probability of occurrence of certain events (inputs, procedures, outputs) is high. In indeterminate systems, the probability is lower. Where the probability is low, the operator must make choices among responses; hence, indeterminate systems require the use of complex decision-making processes and determinate ones do not. (p. 15)

In measuring man-machine system performance, we are concerned with the processes by which inputs to the system are transformed into stimuli to the operators of the system, and, after some form of processing, are transformed first into operator outputs and then into system outputs.

When input uncertainty is high, the system possesses a large amount of indeterminacy. It should be noted that a system may be highly indeterminate even at the lowest levels in its hierarchy. For example, a sonar operator may be called upon to make judgments concerning the source of an uncertain target signal. These judgments become inputs to tactical decision making at all higher levels in the system hierarchy, but they involve a great deal of uncertainty. As a consequence, decisions at the higher levels cannot be fully understood, much less evaluated, unless the uncertainty at each lower level can be identified. Since uncertainties in a system may have many sources, the evaluation of system output must take as many of these sources into account as possible.

THE ELEMENTS OF MAN-MACHINE PERFORMANCE MEASUREMENT

Because it involves the behavior of people, man-machine performance measurement is subject to the same complexities and constraints regarding the drawing of defensible conclusions as are psychological experiments. There are problems of experimental control, of properly handling all operationally significant variables, of the realism of inputs and their proper description, and of measurement of the many categories of response throughout the system hierarchy. There are also the issues of replication of results, repeated trials, representativeness of subjects, and, in general, experimental designs that are adequate to answer the questions being asked. Since man-machine performance measurement can be quite costly, it seems inescapable that compromises between practicality and experimental elegance will have to be made.

However those compromises are made, the measurement of man-machine system performance includes at least the following elements:

1. The definition of what one wishes to know as a result of performance measurement.

2. The identification of appropriate operational criteria.

3. The design of a comprehensive test scenario.

4. The selection of the best test environment.

5. The selection and training of test personnel.

6. The development of data collection methodology.

7. The analysis and synthesis of data into operationally meaningful evaluations.

These steps are highly interdependent and each is firmly related to the overall system objectives (Figure 1).

Identifying What One Wishes to Learn

We started by noting the goal-oriented nature of man-machine systems. Before suitable criterion measures of man-machine performance can be selected, it is evident that one must be clear about what system goals are the object of concern; that is, what questions the performance measurement is expected to answer. As obvious as this sounds, investigators do not always clearly identify these questions.

Meister (1976) distinguishes between system, mission, and personnel performance criteria, reflecting quite different requirements for man-machine performance measurement.

> System-descriptive criteria include reliability, maintainability, survivability, vulnerability, cost, acceptability, effectiveness (output), and efficiency (output divided by cost). Mission-descriptive criteria include output quantity and accuracy, reaction time, and queues and delays. Personnel performance criteria are associated with individual operators and crew responses: reaction time, accuracy, and response variability (pp. 12-13).

In the measurement of man-machine systems performance, we may to some degree be concerned with all three of the types of system goals implied by these criteria. Since systems are hierarchical, we must be concerned with the performance of individual operators throughout the hierarchy. We must eventually be able to relate performance at each level to broader mission criteria. Since this paper focuses on man-machine performance, however, we will be less concerned with such factors as system reliability and maintainability because, despite their obvious relevance, they are not a direct reflection of performance per se. The same can be said for criteria such as vulnerability; however, we may well be concerned with the operability of the system in the event of casualties to either personnel or equipment (i.e., ability to operate in a degraded mode).

Meister has further noted that, because these three classes of criteria define three different aspects of systems, the investigator may well obtain different evaluation outcomes, depending upon which type he uses. In fact, much of the difficulty in interpreting behavioral effects, he feels, results from the investigator's use of multiple criteria. However, he is quick to point out that a single criterion may produce an erroneous picture of system performance, and concludes that at least two types of measures (individual and system) are required to describe the system meaningfully; neither measure alone can supply adequate data.

However one feels about this argument, emphasis in man-machine performance measurement clearly must be upon total system operation as opposed to the partial task performance sometimes examined in the more limited context of skill evaluation. It was noted earlier, however, that total system output often cannot be understood if it is not relatable to the performance of individual operators within the system. In this sense, there is little distinction, I believe, between man-machine performance measurement and team performance measurement. Both imply the ability to identify and assess performance at all levels in the system (team) hierarchy. Glaser, Glanzer, and Morten (1955), in performing detailed job analyses of the functioning of Navy CIC teams, found it desirable to break down every act of the team members into three elements: input, the signals or stimuli that elicit the behavior; process, the response; and output, the signal or stimuli resulting from the process. It was noted that the output can usually be linked to the next act performed by the team since the output of one member usually becomes the input of another. This three-stage description of events seems indistinguishable from the functioning of man-machine systems. Certainly, the criteria of man-machine system performance must be relatable to input and processing variables, as well as to output throughout the hierarchy.

Identification of Appropriate Operational Criteria

The question of what one wishes to learn from man-machine performance measurement heavily influences the selection of operational criteria. In general, global measures of operational (combat) capability are not diagnostic of subsystem performance capabilities and limitations. Subordinate criteria will almost certainly have to be identified if the system output is to be related to man-machine performance at lower levels in the hierarchy, upon which it is certainly dependent.

It should be noted that criteria that are useful for engineering tests are not necessarily meaningful for operational readiness tests. For instance, a commonly used criterion of sonar system performance, the recognition differential, is defined as that level of signal strength that produces, with an alerted operator in the loop, a probability of detection of 0.5 for some specified false alarm rate. Such a criterion may be useful for comparing the detection sensitivity of System A with that of System B. For a test of more general detection performance capability, however, one might be more concerned with the signal strength required for a higher probability of detection under nonalerted conditions. If one infers, from the value of the recognition differential, some corresponding level of operational performance under routine watchstanding conditions, there is a danger of seriously overestimating that level.

In weapons system trainers, instructor personnel often report that they are able to "sense" the proficiency level of the personnel whose performance they observe. It is obvious, however, that in most complex weapons systems it is a practical

impossibility for an observer to even observe, much less record, the multitude of inputs and outputs of the various man-machine subsystems or to determine how the performance of one component of a subsystem affects the performance of the others and the total system output. The typical procedure in postexercise analysis is to focus on such global criteria as how many friendly units were lost to enemy action, how many missiles were fired at the enemy and with what results, how effective the search or intercept plan was, and so forth. While these may be regarded as some of the ultimate criteria of interest, they provide little if any information about the strengths and weaknesses of subsystem components, about problems of team coordination, or even why a particular operation was or was not militarily successful. In fact, it has long been recognized that the hit/miss criterion of combat team performance is a very unreliable one; a hit may result despite very poor performance in some subsystem elements, or a miss may occur because of a variety of uncontrollable factors despite excellent, well-coordinated team performance.

To fully understand the performance of a man-machine system at all levels in its hierarchy, it thus becomes necessary to identify a corresponding hierarchy of performance criterion variables. These can range from the most elementary type of action engaged in by a single operator in the system (speed of button pushing) to the very global measures of team output which, in the extreme, may simply reflect whether the attacking unit or the attacked unit was the survivor. An attempt to convey an impression of such a hierarchy is reflected in Figure 2 which we somewhat arbitrarily have divided into three levels using classical descriptors: proximate criteria (Level 1); intermediate criteria (here shown as Levels 2, 3, and 4); and ultimate criteria (Level 5).

The examples in this illustration relate to the Navy Tactical Data System (NTDS). They do not reflect a full analysis of NTDC operations but are sufficiently exhaustive to convey the concept of a hierarchy of criteria. Level 1 includes many of the basic operator and user activities essential to NTDS functioning. The relationship of some of the measures at this level to eventual mission success is not always obvious. Levels 2, 3, and 4 progressively reflect more and more direct consequences for mission success. In the case of Level 4, the criterion measures begin to involve very consequential decision-making activities. It is probably true that more decision-making is associated with the higher levels of a system; that is, there is more indeterminacy there but, as pointed out earlier, very consequential decisions can occur at the lowest levels as well. Finally, Level 5 includes a variety of criterion measures that essentially reflect the outcome of the engagement. These are the ultimate criteria that are usually employed in postexercise evaluations.

The connected items in Figure 2 trace one example of how performance criteria at different levels in the system might relate to one another. At Level 1, the number of unnecessary actions, such as operating mode changes made by a console operator, may affect the time taken (at Level 2) to enter a new track which, in turn, influences the time taken for the target to be classified (Level 3), which influences the decision to launch an interceptor and thus the intercept attack time (Level 4) which, finally, affects the distance from the target's weapon release point at which the target is engaged (Level 5).

It will be noted that many of the performance criteria at Level 1 simply have to do with the time taken to process system inputs and to transmit an output to the next level. However, it is clear that accuracy of response is also quite important.

In fact, timeliness and accuracy of response are pervasive dimensions of performance criteria at all levels in the hierarchy. At higher levels in the system, corrective actions also appear as important performance criteria. More about this criterion measure later.

An important thing to note about the figure is that all of these criteria, at least for a system like NTDS, are potentially measurable. In spite of this, some of the most significant behavioral criteria are not identified. This is the behavior that has to do with the human (or machine) processes that intervene between input and output. These are often the largest sources of variability insofar as the human element of a system is concerned and, in many man-machine system contexts, may be the most consequential criteria of all.

Design of the Test Scenario

It is clear that the outcome of any measurement of man-machine performance is largely determined by the characteristics of the test scenario. If the purpose is to determine whether system personnel are capable of performing combat tasks (i.e., their operational readiness), the characteristics of the mission scenario are critically important, particularly with respect to such variables as information load, complexity of responses, available decision time, system degradation due to environmental factors, and so forth. Since operators respond to stimuli, not raw system inputs, complete realism in the stimulus presentation is usually absolutely essential if a meaningful appraisal of system capability is to be achieved. This is an often neglected aspect of man-machine performance evaluations, particularly of tests conducted in system simulators. The variables that affect system inputs (i.e., signal features) may affect man and machine quite differently. For example, there is no question but that a machine is a far better detector than man of very low intensity signals in noise, provided the signals are steady state. Conversely, man may be a much better detector of transients or of signals whose temporal and spatial characteristics are highly variable or unpredictable.

The types of variables most likely to be included in a test scenario are the characteristics of friendly and adversary weapon systems, the type and number of targets that are encountered and where they will be encountered, and various environmental factors that may affect tactical alternatives.

Among the many variables that can markedly influence man-machine performance but which often are not appropriately reflected in the test scenario, especially when the test is conducted in a simulator, are:

1. Information load (note that man-machine systems performance may degrade under either very high or very low information loads).

2. Task duration (experimental task durations are often far shorter than real-life operational tasks).

3. Degree of operator alertedness (usually much higher than in routine operations).

4. As noted earlier, various aspects of stimulus (input) realism that may markedly affect such critical behaviors as probability of target detection, speed of localization, and accuracy of classification.

In addition to its problem content, another highly important characteristic of the test scenario is its provision for gathering reliable performance measures. In this respect it is quite like other tests involving human behavior whose reliability heavily depends on multiple samples and/or repeated observations. In his excellent review of team performance, Glanzer (1962) emphasized that team training programs are often least satisfactory in the manner of proficiency measurement because, when team proficiency measures are used, they tend to be restricted to one or very few problems. If the performance test has relatively few "items," it is also likely to have low reliability. One technique suggested by Glanzer to compensate for the single-item character of team performance tests is multiple scoring of individuals within the team. Since each team member usually carries out several acts, it is possible to derive a score based on several acts for each individual. This, he felt, would be an important supportive technique in measuring the team's efficiency. Because it is possible to have the majority of individual members of a team do well and yet have the team function poorly as a unit, this technique might focus on the possible causes of poor team performance.

Also related to performance test reliability is the problem of changes in performance as a function of time. Meister (1976) has noted that the responses of operator personnel change over time as a consequence of input changes, practice, fatigue, and motivational variability. In contrast to the typically very slow changes that occur in equipment characteristics and outputs, which may require thousands of operational cycles to manifest themselves, changes in operator behavior occur quite rapidly, often within a signle mission or operational cycle. Such changes, generally speaking, are a source of unwanted variance and, therefore, unreliability in the performance test results, although for some evaluative purposes they may be of interest in themselves.

Finally, it should be noted that the reliability of performance tests can be enhanced by incorporating into the test scenario several opportunities for each team member to perform various critical operational tasks under similar, but not identical, circumstances. If several such performances are separately measured, then the results, as with all behavioral testing, are more likely to be reliable. This approach is more feasible, of course, if the test occurs in a system simulator than if it occurs in the operational environment. Partially repeatable problems are achievable in simulators by clever programming of the system inputs. For example, it can be arranged so that no matter what tactical actions are taken by the team, a particular target type is encountered at the desired time in the scenario, at the desired bearing and range, and makes the desired maneuvers. Such elements of problem control, and thus repeatability, are obviously difficult if not impossible to achieve in the operational environment. This leads to the next consideration, the choice of test environment.

Selection of the Test Environment

Perhaps the most significant choice facing the designer of a man-machine performance test is that of whether the test will be conducted in the operational environment or, if one exists, in a suitable operational simulator. The real environment is obviously the one of greatest apparent operational relevance; for a measurement point of view, however, the simulator environment can provide a multitude of advantages.

The possibilities for inclusion of all relevant variables in the test scenario are closely related to the selection of the test environment. There appear to be three basic alternatives:

1. The operational environment is used and the problem scenario is specified in general, but all encounters, system inputs, and environmental factors are essentially uncontrolled. This is the approach most frequently used for operational tests and evaluations.

2. The mixed operational/simulation approach in which one or more elements of the problem scenario are artifically controlled, but the operating environment and equipment are real. An example is the injection of artificial or recorded target signals with known characteristics at specified times into an otherwise normally operating system.

3. Full-scale simulation in which the inputs to the system, the equipment, and the operating environment are specifiable, controllable, and have various degrees of realism. The equipment may or may not be functionally identical to the operational system, although an attempt usually is made for the man-machine interface to be so.

Each of the above choices usually leaves the investigator with some methodological problems. Perhaps the greatest advantage of the operational environment is its unchallenged "realism." However, it must be recognized that any measurement conducted in the operational environment is likely to be contaminated by a multitude of uncontrolled variables. Consequently, the operational test may or may not be representative of the operational capability of the system since available resources usually do not permit the conduct of repeated observations under different environmental conditions. In other words, the outcome may very well be unreliable. In contrast, repeated measures of performance in the system simulator are often readily achievable; the problem here, of course, is that not all influential real-world variables may be effectively simulated or even identified.

Among the problems often associated with man-machine performance measurement in the operational environment is the occurrence of various nonprogrammed and even unknown inputs, the lack of opportunity for task replication, the fact that systems often operate in a degraded mode, the possible requirement to measure performance on a not-to-interfere basis, and the fact that the system personnel may be less under the control of the test officer than in a system simulator.

In terms of the need for task replication, it must be noted that in a complex man-machine system, operator interdependencies may make it very difficult to achieve problem standardization. It may very well be a practical impossibility to specify all of the inputs to the operators and therefore to assess their outputs. Yet some replication (within limits) is very necessary for a reliable assessment of man-machine system capability and there seems little doubt that it can be more readily approached in a simulator than in the operating environment.

On the other hand, it should be noted that many environmental and operational stresses that may be very important to real-world operations are rarely represented in a simulator. Among the more obvious ones are motion stress, heat, noise, vibration, acceleration, atmospheric contamination, extended periods of sleep loss, boredom, monotony stress, and so on. Obviously, any of these variables may markedly affect man-machine performance.

In addition, when system simulators are used for performance evaluation, a frequent complaint of military personnel is that the simulation is "too perfect" compared to the presentation typically experienced in the operating environment, particularly inasmuch as operational equipment often is not found in peak operating condition. The central issue, of course, is whether the system inputs, on which everything else depends in either environment, are an appropriate representation of the operational problem considering the objectives of the man-machine performance test.

Finally, whichever environment is used, the experimental procedure should be such as to reduce the "test" characteristics of the test as much as possible. The special motivating influence of "test" conditions, perceived as such, requires little comment. Yet their effects are often overlooked. The unobtrusive injection of controlled test signals into the otherwise routine operating environment may be a particularly effective means for achieving this objective.

Selection and Training of Test Personnel

Man-machine performance tests are often conducted as if the tested personnel, granted that they hold appropriate technical designators, are essentially homogeneous in operating skills. In reality, in many military systems there are large individual differences in operating skills within a given pay grade and specialty class. Individual differences in skill level are likely to grow, not to diminish, as systems personnel advance in their careers (Figure 3). As a consequence, uncontrolled individual differences in personnel skill levels may have more to do with the outcome of a man-machine performance test than the other system variables under study.

Related to this problem is the fundamental issue of whether the test designer wishes to perform the test with "typical" or superior personnel. Clearly, this depends on the objectives of the test. The problem, however, is that there is a tendency for operational tests to be conducted using personnel who are superior performers, either by virtue of their past operational experience, or because they were given more than routine training in system operation prior to the test, or both. Thus the man-machine test may well produce an erroneous impression of how well operational systems like the one under assessment will perform when more representative personnel and more routine training are employed.

Glanzer (1962) has pointed out the very large role played by individual proficiency in team performance. He emphasized that "Skilled activity by an individual team member means that the individual's responses meet certain requirements of timing, coordination, and sensitivity to changes in the environmental situation. Skilled activity by a team means meeting the same requirements."

In a study of the errors made by CIC teams during ship control and gunnery exercises, Glanzer and Glaser (1955a) found that error rates were somewhat higher for high-ranking team members than for the lower-ranking ones and that only a small proportion of the errors were corrected within the team. They attributed this possibly surprising result to the fact that in most cases the responsibility for correction was not clear. From this and earlier studies, they concluded that the two principal difficulties characterizing poor team performance were: (1) errors committed by individual team members, and (2) inefficiency in correcting these errors. They felt that, although more complex factors such as

"coordination of personnel" might play a role, they were much less prominent sources of variance in team performance than were the errors of individual team members and the failure of any member of the team to correct them. From this, and from an abundance of experimental evidence, we conclude that individual differences in operating proficiency are very likely a major source of variance in man-machine systems performance as well.

Data Collection Methodology

Historically, procedures for collecting man-machine performance data during operational tests have been fraught with difficulties: desirable recording procedures are likely to interfere with operations, many significant processes are likely not to be recorded, and data reduction and exercise reconstruction are often a tedious, time-consuming process.

Recently, significant progress has been made toward overcoming many of these problems, particularly in computer-centered systems. Perhaps the best known and most completely developed approach is the Operational Performance Recording and Evaluation Data System (OPREDS). As used in a system such as NTDS, OPREDS records, on a common time-base, all inputs to the computer from the system consoles and all computer outputs back to those consoles. For example, it provides a continuous on-line record of all operator interactions with the system computer, together with the identification of the x-y coordinates of all local and system tracks, and the adjustments made by all operators in updating all tracks. Data for the fire control system can be captured, beginning from weapon assignment to time of directed search, lock-on, missile deployment, and splash.

The data reduction program developed for OPREDS will provide a plot of any and all target tracks known to the NTDS, along with the concurrent time history of all sequences of team member actions, in an operational sequence diagram format. In addition, an auxiliary system has been developed that permits the recording of all voice commentary on the same time line so that, in postanalysis, voice communications occurring at any particular times of interest can be quickly selected out and analyzed for content.

The OPREDS system is totally noninterfering with the operational system's hardware, software, and personnel. The recording unit is easily portable, installable in minutes, and requires only one operator-observer. Recent improvements include an on-board analysis capability with a time delay of only 1 day following an operational exercise (Urmston, 1977).

Although OPREDS was developed specifically for performance recording with NTDS, the concept clearly should be generalizable to any system having a central computer through which all system inputs and outputs flow. The necessity of data recording devices such as OPREDS for monitoring and recording the performance of any relatively complex man-machine system is obvious. Clearly there are severe constraints on how much information any test observer or even group of observers can monitor; OPREDS makes it possible to automatically and objectively measure the performance output of individual team members as well as identify the sequential actions and interactions of various members of the team. It also promises the opportunity to trace the sources of errors and delays in the system, and to determine the impact of these on total system performance.

Exercise reconstruction rarely pinpoints the reasons at the operator level for good or poor system performance. Usually, neither the critical stimuli nor the critical responses can be identified. With the development of recording systems like OPREDS, a great deal of progress can be made toward recording these aspects of man-machine functioning. Some typical summary data available from OPREDS are shown in Figure 4. It will be evident that OPREDS does not yet directly yield evaluative performance measures, although it does provide some extremely useful beginnings.

Performance Evaluation

There are numerous methodological problems associated with the evaluation of man-machine performance data. The most fundamental is that, in attempting to assess man-machine system performance in the operational environment, the maximum achievable performance is often difficult to specify because of the unknown effects of a wide variety of uncontrolled variables. Thus it may be very difficult to set a standard of how well the system could have performed. This is one of the considerations that makes the use of system simulators attractive for the measurement of man-machine system performance. In the simulator, it may be possible to define the performance of a theoretically perfect man-machine system because appropriate responses to known system inputs can be much more fully specified.

In the simulator the scenario can be arranged so that the test director knows "ground truth." Some knowledge of ground truth, that is, the true state of affairs at the times man-machine system performance are measured, is fundamental to the assessment of performance. In a test conducted in the operational environment, an attempt to establish ground truth is usually made through knowledge of the general scenario design and exercise reconstruction. However, considerable uncertainty may be associated with this process, even if the total exercise is highly constrained, which often it is not. For this reason, controlled signal injection into an actual operational system may be desirable. When signal injection is used, at least a part of the input to the system can be specified exactly. Thus the output, which represents the combined result of the machine and man operating on that input, can be assessed with far greater certainty.

The ability to adequately describe the operational environment is also a part of "ground truth." While the real-world environment can be described in general terms, and its effects predicted to some extent on the basis of theory, the simulator affords the opportunity to vary these effects much more conveniently and to specify them exactly. However, as noted earlier, many important environmental variables are absent from simulated environments.

Another major problem associated with the evaluation of man-machine performance data is that the operational scenario is often open-ended in the sense that, once a sequence of actions begins, subsequent inputs to the scenario are partially determined by the uncontrolled reactions of the various team members. As noted earlier, some degree of control can be maintained in the operational setting through signal injection, or in the simulator through more extensive manipulation of the entire scenario. But the typically employed "free play" exercise, whether in the operational environment or in a simulator, places the evaluation of man-machine performance, except in terms of the most global systems criteria, almost beyond reach.

Finally, certain other aspects of man-machine performance evaluation remain particularly vexing. These include assessing the effect of operator interdependencies (coordination, anticipation) on system output and evaluating outputs that are made in verbal form. (It should be noted that much of the output of many man-machine systems is verbal.) Some progress on both of these points has been made possible through systems such as OPREDS. But more fundamentally, as noted earlier, many of the critical processing behaviors of operators are not easily related to their overt outputs. This is, often the system output can only be examined in terms of superficial operator responses such as button-pushing. Much of the decision making that takes place in a man-machine system thus cannot be directly observed or measured and, as noted earlier, indeterminate systems involve many uncertainties that require decisions at all subsystem levels. The measurement of decision-making behavior remains a challenge whether that behavior occurs in the cerebral cortex of a man or in the algorithm of a computer. It is in this direction, perhaps, that future research and development work should move.

Summary and Conclusions

1. Man-machine systems are goal-oriented, hierarchical in nature, and vary in their degree of determinacy. Each of these characteristics influences how man-machine performance should be measured.

2. The principal purposes of man-machine performance measurement are to assess operational readiness and to diagnose sources of deficiency in readiness. Deficiencies may be associated with system hardware, software, or personnel, or with the interactions among combinations of these subsystems.

3. Deficiencies in total system output often cannot be understood and corrected unless they are considered in relation to performance at appropriate subordinate levels in the system hierarchy.

4. A full description of man-machine system performance requires the specification of inputs, processing, and output. In many system performance tests, much data are available on outputs, somewhat less on inputs, and little, if any, on processing. Understanding the output, and making correct inferences about the intervening processes, are heavily dependent on how adequately the inputs are described.

5. Man-machine performance criteria must reflect broad system goals. However, there is a hierarchy of performance criteria corresponding to different levels of the system hierarchy that may be quite specific in nature. Global criteria, such as those typically employed in postexercise reconstruction, are usually very uninformative about man-machine performance at the subsystem level.

6. In a test of man-machine performance, the test scenario must be designed to reflect all operational, environmental, and personnel variables that impact on the ability of the system to achieve its goals. Some variables affecting man-machine performance that are often neglected, especially when measured in the system simulator, include operator skill levels, information load (either high or low), task duration, personnel alertedness, and equipment degradation.

7. Man-machine performance tests involve all the subtleties of carefully designed psychological experiments. However, practical constraints nearly always force serious compromises with respect to the experimental design, particularly when the test is performed in the operational environment.

8. The likelihood of obtaining reliable man-machine performance measures is a function of the number of measurement opportunities and problem replication. In the operational environment, full replication of test conditions is probably never achievable; in the simulator environment, it can be approximated through manipulation of the scenario.

9. In the simulator environment, the effects of many operationally relevant stressors are usually absent. These include motion stress, acceleration, heat, noise, vibration, atmospheric contamination, sleep loss, boredom, and monotony stress.

10. The input variables in simulators used for measuring man-machine performance are often simplified to the point where many of the complexities encountered in the operational environment do not arise. Very misleading impressions of performance capability can result.

11. Mixed operational/simulation test environments may represent an optimal compromise between the need for experimental control, ability to specify the system inputs, and realistic effects of environmental variables.

12. The practice of using superior or specially trained personnel for man-machine performance tests can be a source of misleading information concerning more typical man-machine system performance.

13. Ability to specify "ground truth" is highly important to the evaluation of the performance of many man-machine systems. This is very difficult to do with "free play" test scenarios, regardless of the test environment.

14. Recent advances in unobtrusive data recording systems have made it possible to capture much more detail on man-machine subsystem performance than heretofore. Nevertheless, much significant processing behavior that may be important, particularly in nondeterminant systems, is not captured.

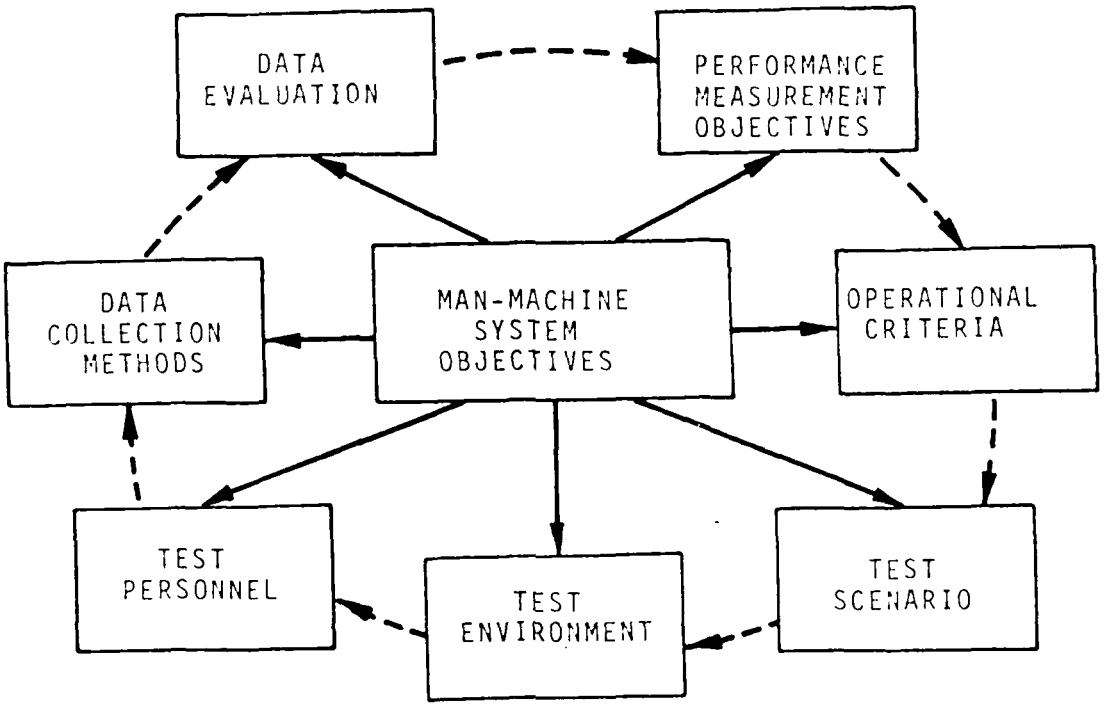


Figure 1. Interdependent nature of the elements of man-machine performance measurement

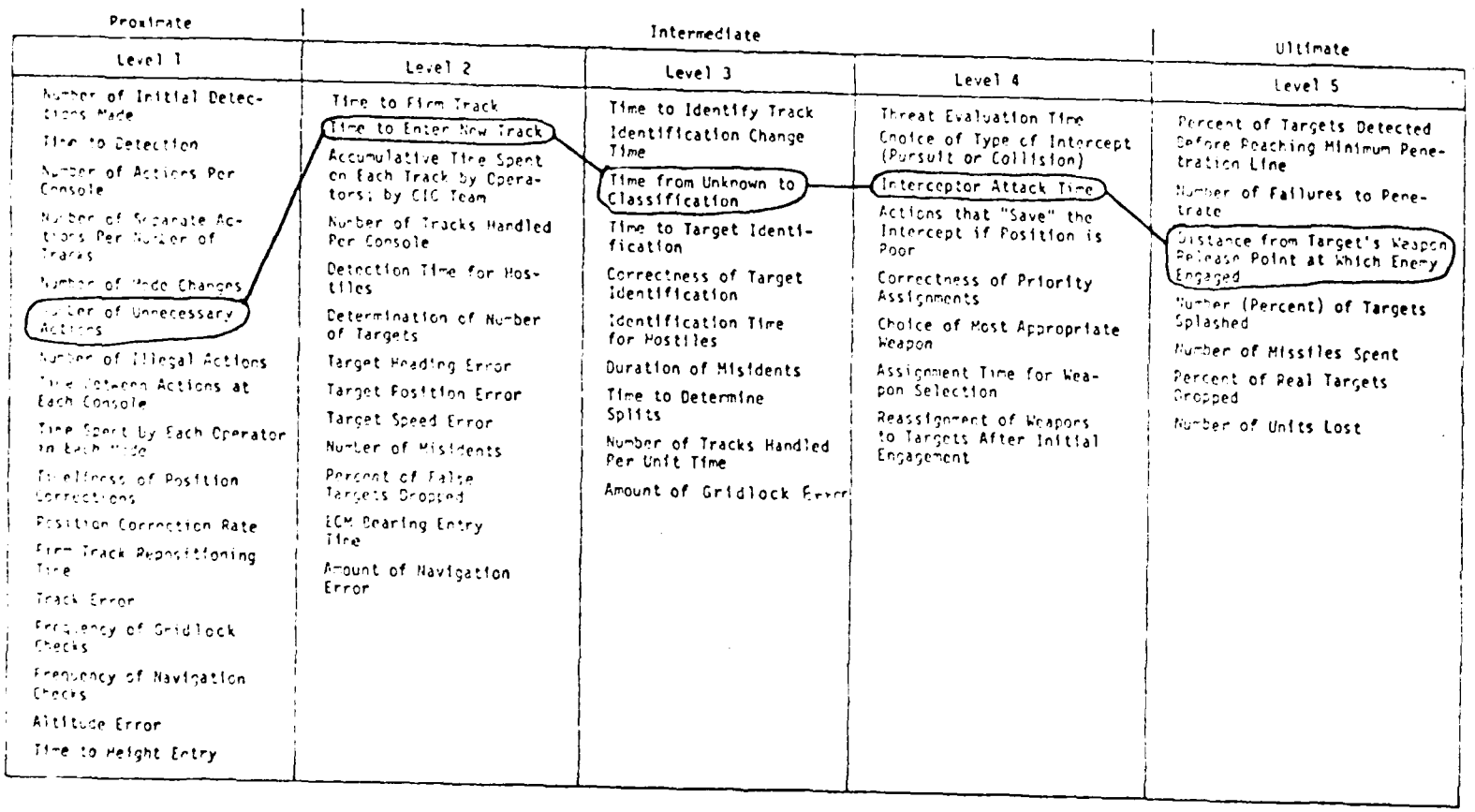


Figure 2. Levels of performance criteria.

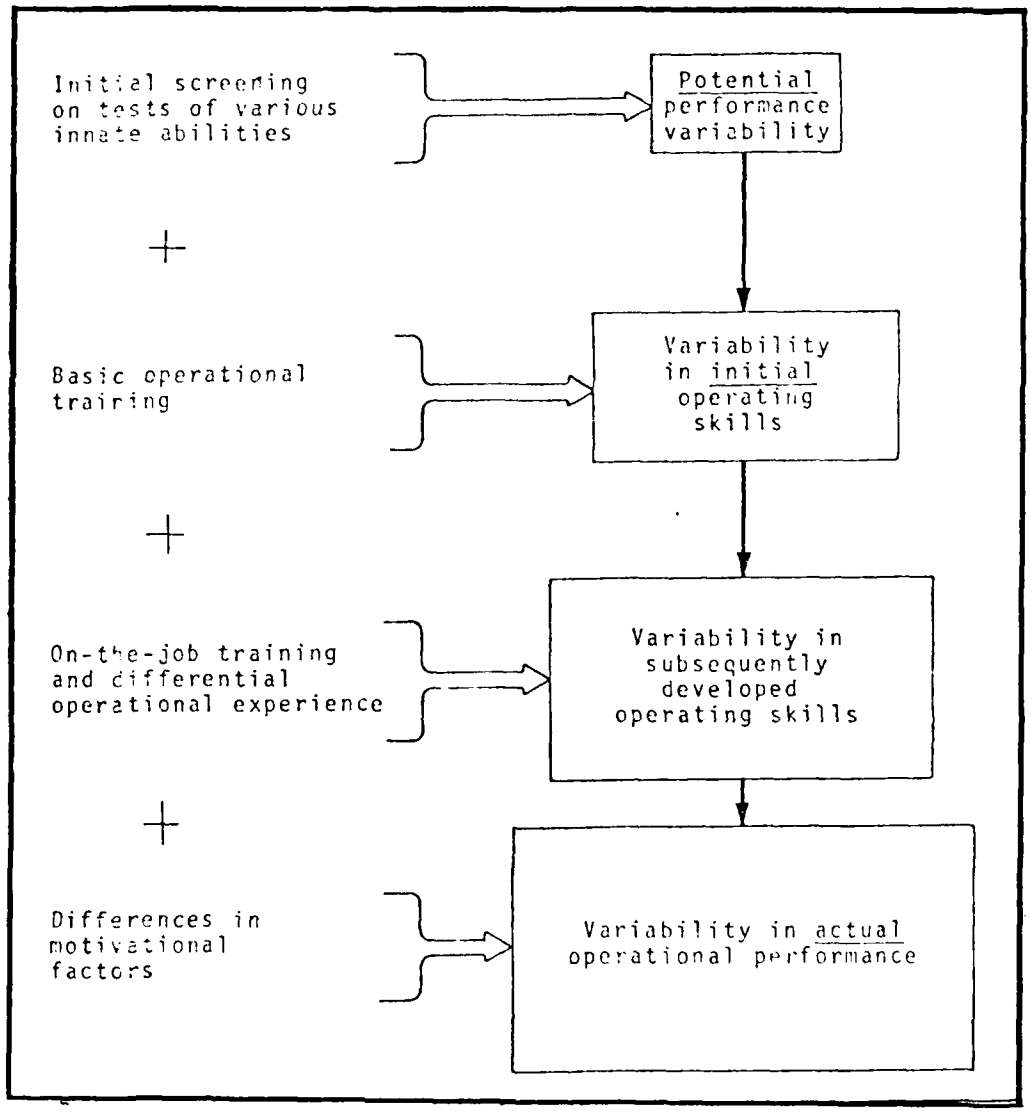


Figure 3. Increasing performance variability as a function of innate abilities, training, operational experience, and motivation.

- Number of separate actions in the system as a function of time
- Number of tracks handled
- Number of actions per console
- Number of tracks per console
- Time between each action at each console
- Time between sets of events
- Time from unknowns to classification
- Accumulative time spent on each track by operators
- Accumulative time spent on each track by CIC team
- Total time spent by operator in each mode
- Time to firm track
- Time to identify track
- Time to weapon assignment
- Time to assign missiles
- Total engagement time

Figure 4. Some available outputs from OPREDS.

REFERENCES

Chapanis, A. Research techniques in human engineering. Baltimore: The Johns Hopkins Press, 1959.

Eckman, D. P. (Ed.) Systems: Research and design. New York: John Wiley and Sons, 1961.

Glanzer, M. Experimental study of team training and team functioning. In R. Glaser (Ed.), Training research and education. Pittsburgh, PA: University of Pittsburgh Press, 1962.

Glanzer, M. and Glaser, R. Performance characteristics of three types of Navy teams (ONR Technical Report). Washington, DC: American Institute for Research, 1955a.

Glanzer, M. and Glaser, R. A review of team training problems (ONR Technical Report). Washington, DC: American Institute for Research, 1955b.

Glaser, R., Glanzer, M., and Morten, A. W., Jr. A study of some dimensions of team performance (ONR Technical Report). Washington, DC: American Institute for Research, 1955.

Klaus, D. J. and Glaser, R. Increasing team proficiency through training: 5. Team learning as a function of member learning characteristics and practice conditions (No. AIR-E1-4/65-TR). Washington, DC: American Institute for Research, 1965.

Meister, D. Behavioral foundations of system development. New York: John Wiley and Sons, 1976.

Parsons, H. M. Man-machine system experiments. Baltimore: The Johns Hopkins Press, 1972.

Sheridan, T. B. and Ferrell, W. R. Man-machine systems: Information, control, and decision models of human performance. Cambridge, MA: Massachusetts Institute of Technology, 1974.

Wylie, C. D., Dick, R. A., and Mackie, R. R. Toward a methodology for man-machine function allocation in the automation of surveillance systems. Vol. 1, summary (Technical Report 1722-F). Goleta, CA: Human Factors Research, Inc., 1975.

Urmston, R., Naval Ocean Systems Center, San Diego, personal communication, 1977.

ABOUT THE AUTHOR

Dr. Robert R. Mackie is President and Director of Research at Human Factors Research, Inc. Dr. Mackie received his Ph.D. in quantitative psychology from the University of Southern California in 1950 and has been engaged in applied experimental research on human behavior for the past 27 years, and has published over 80 technical papers in his field. Directing an interdisciplinary team of psychologists, engineers, physiologists and physical scientists, his work encompasses a wide variety of problem areas: perception of complex stimuli; judgment and decision-making; learning and training; attitude formation and change; attention and alertness; performance under environmental stress; man-machine interaction in complex systems; highway safety; and techniques for improving the research-to-application process in the behavioral sciences.

PERFORMANCE MEASUREMENT TECHNOLOGY: ISSUES AND ANSWERS

Earl A. Alluisi
University Professor of Psychology, Performance Assessment Laboratory
Old Dominion University, Norfolk, Virginia

ABSTRACT

The trend over the past decade, which has led to greater emphasis on performance assessment technology, is briefly reviewed, and the characteristics of performance assessment research are discussed with special attention to the criterion problem and the alternative methodologies based on job-performance, simulation, synthetic-work, and specific-test techniques. Five future issues for performance assessment research and development are presented and discussed. These are: (1) the convergence of this technology with those of selection, training, and job design to provide for non-discriminatory personnel practices as required by law; (2) increased attention to crew, group, team, and unit performance assessments as differentiated from individual performances; (3) the development of operator-system transfer functions in order to relate measurements of system performance with performances of individual human operators; (4) the need to attend to questions of optimum (or even appropriate) degrees and kinds of fidelity in simulation for different usages; and (5) the application of performance assessment technology in measuring and predicting operator workloads and in the specification of optimum loading levels.

As Ben Morgan and I observed last year in our Annual Review of Psychology chapter on "Engineering Psychology and Human Performance," the trend over the past decade has been toward ever-broadening applications of human performance research and performance measurement technology in "the design, maintenance, operation, and improvement of all kinds of operating systems in which humans are components" (Alluisi and Morgan, 1976). This symposium on "Productivity Enhancement: Personnel Performance Assessment in Navy Systems" is yet another demonstration of the continuation of that trend. Increased emphasis on performance assessments (i.e., performance measurements and evaluations) is an important constituent of the trend, and it might be of some use in predicting the future to review briefly the recent past in this area. By doing so, we shall be able to refresh our memories on the forces, findings, notions, and needs from several different directions that have converged to provide a new emphasis and direction to performance assessment technology.

TREND TOWARDS PERFORMANCE ASSESSMENT TECHNOLOGY

Performance Measurement Methodology: Relevance

Ten to 15 years ago, the relevance of the then-current kinds of laboratory research on human performance was being questioned. Even in the scientific community, doubt was raised regarding the capability of the findings of such research to be generalized and implemented in practical situations (cf. Chapanis, 1967). Today, these

PRECEDING PAGE BLANK-NOT FILMED

are no longer pressing issues, and the questions seem not to be necessary except in rare instances of highly "academic" research. Instead, the pressures today are to provide cost-effectiveness analyses of the applications of research and development findings, and even to employ analyses of potential cost effectiveness benefits as part of the criteria for the allocation of resources among different, alternative, and often competing research and development programs, projects, tasks, and work units.

In short, "relevance" and potential applicability have come to be accepted as absolute requirements, and now the trend indicates a need to develop methods for making estimates of potentials so that cost-effectiveness potentials can be employed as a strategy in deciding among alternative research and development efforts. Although this push has doubtless made some differences in the research and development that is being done, it would be a mistake to conclude that all of the efforts have been or are going to be changed. Certainly, it can be assumed that some of the efforts were "cost-effective," even if unanalyzed, and that, in these cases, it has been merely the demonstration of potentials in the language of managers that has been added to the repertoire of the performance-assessment researcher. We are following the engineers in this, and we certainly have a lot yet to learn by way of analytical methods for estimating cost-effectiveness criteria not only for the selection of which research and development alternatives should be supported and to what extent, but also to provide the necessary leverage for application and implementation of findings and for demonstration of system life-cycle cost-effectiveness alternatives. It should be clear that, as these methods are more fully developed and as the potential (and actual) benefits of "people-oriented" research and development are able to be expressed in the same management-relevant language that has been used by the "hardware-oriented" researcher for nearly 20 years, opportunities will be increased for possible shifts in the levels of support and total efforts in the two areas. Performance-assessment research and development will increase to the extent of its capability of demonstrating its worth in cost-effectiveness terms. This assertion is not based on any assumption that requires a logically consistent management system, but rather on the empirically demonstrable observation that the behavior of such systems is shaped and maintained by the contingent probabilities of the applicable reinforcements. Success breeds success, especially in the research and development arena. Indeed, the promise of success increases the probability of support, and increased support increases the probability of success.

Thus, far from being fearful of the trend, I am extremely optimistic. I believe that we have only to learn and apply (perhaps after adapting and extending) some relevant methods from econometrics in order to demonstrate the real and the potential benefits of performance assessment technology and thereby to "win" more in the budget-allocation "battles" that must always be fought.

Performance Measurement Methodology: Content

An international symposium was held in Amsterdam during September 1969 at the instigation of Professor A. Chapanis and under the sponsorship of the International Ergonomics Association. The presentations made during the symposium served as the basis for a text on the Measurement of Man at Work (Singleton, Fox, and Whitfield, 1971). The majority of the 27 papers dealt with performance measurement methodology and the techniques of measuring man at work both in the laboratory and in the field. The European work tended to be oriented towards the use of

psychophysiological criterion measures (e.g., heart rate and energy costs), whereas the American work tended to be oriented towards the use of behavioral measures (with the British between the two). More importantly, the emphasis of the studies reported tended to be on atypical rather than the more usual tasks and jobs--e.g., piloting rather than driving, stress rather than normal conditions of work and performance, and specialized rather than general populations of workers.

At about the same time, a bit more than a decade ago, a conference at the Aerospace Medical Research Laboratories led to an American Psychological Association symposium on "Methodology in the Assessment of Complex Performance" and the publication of seven papers in an issue of Human Factors specially edited by W. Dean Chiles (1967). These papers covered questions of methodology and measurement in field research, full-scale mission simulation, factor-analysis-based tests assembled as a battery in a single apparatus, and the synthetic-work methodology which also used a battery of tasks, but with the requirement for time-sharing and multiple-task performances. Among the more important problems of performance assessment methodology identified were those of (1) the criterion problem, (2) the taxonomy(ies) of tasks, (3) the reliability of performance measures, and (4) the role of face validity, especially with regard to the subject's approach to the work/test situation.

Today, the Europeans are still oriented relatively more towards psychophysiological measures than Americans, probably because psychology is more a captive field of medicine there than here. But both have moved more towards the performance assessment methodologies that involve task-related performance measurements and away from the purely psychophysiological and laboratory or "academic-concept-based" test measurements. The shift is doubtless related to the contingent probabilities; we are finding that we are more often successful in the research and development efforts based on task-related performance measurements, and less frequently successful with the other kinds of measurements.

As part of the same trend, and resulting partly from the greater capability for task-related performance measurements provided by advances in technology such as those that have made possible the newer and more powerful devices like current flight simulators for research and development as well as training, more and more studies are dealing with the typical (rather than the atypical) tasks and jobs. Given that there are so many more persons involved in the "typical" tasks and jobs than there are in the "atypical," and given the movement towards the use of cost-effectiveness criteria in deciding among alternative research and development programs, it is probable that this trend towards more research on the "normal" is going to continue. The potential impact of findings that affect a million workers is much greater than that of findings that affect a few score of workers!

Only one of the four problems of performance assessment methodology cited earlier seems to be of any great importance today--namely, the criterion problem--and we shall discuss that more fully in a subsequent section. After some relatively major efforts directed to the problem of task taxonomies provided advancement, but no clear solution (Fleishman, 1975) human performance researchers seem to have adopted the attitude that it may be more difficult to derive appropriate task taxonomies than to make greater advances directly in the area of performance assessment methodologies and applications. Although still recognized as a problem task taxonomies are much less frequently cited as important problems to be given high priority over other performance assessment problems.

The reliability of performance measures, and their face validity to the subjects, have both taken on less major roles with the technological capabilities to simulate effectively part or all of a given task or job and to measure accurately the subject's performance. This is not to say that the employment of simulation is the answer to all the problems of performance assessment methodology, or even that it is a simple affair; it is neither, and more attention will be given to this topic later.

Performance Measurement Methodology: Impact from Other Domains

Prescience was shown by Uhlaner (1972) in a paper that calls attention to the need for converging the selection, training, and job design areas to develop an optimum methodology for studying the effectiveness of human performance. In fact, no lesser authority than the U. S. Supreme Court has, knowingly or not, mandated such a convergence through several relatively recent rulings based on the Civil Rights Act of 1964. In its well known first equal-employment-opportunity (EEO)-related decision, the Court proscribed the use of employment tests that were discriminatory unless the employer could prove job relatedness. That is to say, if a test employed as a selection device for employment or promotion resulted in lower proportions of minority groups or women being hired or promoted, the test or tests could not be used unless the user could demonstrate its validity in predicting job performance.

Then, in its second EEO-related decision the Court, in the case of the Albemarle Paper Company et al versus Joseph P. Moody et al, addressed the question of what must be demonstrated for an employer to establish that a selection device, racially discriminatory in effect although not in intent, is sufficiently "job related" to conform with the requirements of the Civil Rights Act of 1964. The Court's ruling supported the federal guidelines on employee selection as a proper administrative interpretation of the Act. It said that a test validated on one job cannot be used as a selection device for another job unless, by job analysis or other acceptable methods, it can be demonstrated that there are no significant differences between the two jobs.

Further, since the Albemarle Paper Company had employed an annual performance appraisal or "supervisor rating" as the criterion measure against which the selection device was validated, and since the Court confirmed a broad interpretation of "tests" to include such supervisor ratings, it followed in the ruling that such supervisor ratings could be used properly as criteria only when they themselves met the requirements set for selection (and promotion) devices. That is to say, a vague and general rating (test) could not properly be used; rather, supervisor ratings could be used only where the criteria actually used by the supervisors could be determined and demonstrated to be based on carefully defined job-performance criteria representing behaviors actually required by the job. In short, the Court affirmed or reaffirmed that performance-based criteria were required for acceptable validation of selection devices.

There were other important aspects of the Court's "Albemarle" ruling, but they are not of primary concern to use here so we shall not dwell on them. They had to do with the need for (1) use of equivalent job levels in the validation, (2) minimizing the validation differences between the employee groups that might be studied in obtaining concurrent validity information and the applicant groups whose make up might be quite different otherwise (and to whom the validity information would be applied as "predictive"), (3) conducting separate validation studies

of minority and nonminority groups where feasible, (4) considering alternative nondiscriminatory selection tests or devices, and (5) providing appropriately high levels of professional attention, care, and judgment in the conduct of the validation research. The Albemarle decision was handed down by the Court on 25 June 1975, and its full impact has not yet been felt!

The rulings appear to add up to the following state of affairs: If a woman or a member of one of the minority groups covered by the Civil Rights Act of 1964 can establish that an employment practice adversely affects the class to which that person belongs, then the legal burden of proof shifts to the employer to demonstrate that the standards used for making the decision are job related. Presentation of a rationale, no matter how logical or historically based, is not sufficient to establish the job-relatedness of the practice but, rather, it would be necessary in addition to demonstrate the validity of the practice in terms of job performance, recognizing that supervisor ratings could be used as criteria for such validations only to the extent that they could be demonstrated to be based on carefully defined job-performance criteria. There is little doubt that performance assessment technology, and personnel and training technology, must converge. Some of the implications of these rulings with regard to the military manpower management system and such standard practices regarding selection and classification based on the MOS categorizations will be discussed in a later section.

All told, the requirements of the times have provided an excellent opportunity for major increases in performance-assessment research and development, and for major advancements in the application of performance-assessment technology. The need can be demonstrated, and to the extent that these demonstrations can be translated into analyses of potential benefits (in cost-effectiveness terms), administrative and budgetary support can be increased for this area. It behooves us, then, to consider in greater depth some of the characteristics of the area.

CHARACTERISTICS OF PERFORMANCE ASSESSMENT RESEARCH

Performance assessment is one of the most important and difficult areas of current research; it has been so for some years (cf. Alluisi, 1967). It is at the center of the "criterion problem" for many other areas of research and applications, including the following:

1. The final validation of selection and training techniques depends on the assessments of the performance of the people who have been differentially selected and trained. The importance of these assessments, given the probable impact of the U. S. Supreme Court rulings cited earlier, cannot be overemphasized.

2. The final validation of improvements to man-machine systems by human factors engineering applications depends on the performance assessments of the systems and the systems' operators.

3. The evaluation of the effects of various stresses and the measurement of performance decrements attributable thereto depend on performance assessments.

4. The establishment of optimum operator loads, of operational limits, and even of optimum operational conditions and procedures, and many other tasks depend on the measurement and evaluation of human performance.

Physiologists and engineers, as well as psychologists, have contributed to performance assessment technology over the years. Physiologists have concentrated on those aspects of performance assessment which are easily included within that discipline's expertise; for example, by measuring the output, impairment, and recovery of muscles. Industrial engineers have concentrated on aspects of performance assessment such as time-and-motion study or the measurement of productivity. Psychologists have concentrated their efforts in different ways, reflecting the different subspecialties of psychology: industrial psychologists on training and personnel technology; engineering psychologists on design or redesign of equipment and systems; and experimental psychologists on one or more of the traditional areas of learning, perception, psychomotor performance, etc. For the most part, we can consider these to represent indirect or background research on performance assessment. Direct attacks on the problem, of course, have been made, such as those reported in the special issue of Human Factors 10 years ago and previously cited here (Chiles, 1967), and these have run smack up against two major problems that characterize performance assessment research--the criterion and the R&D strategy. Although these problems are still with us, there is now some glimmer of hope that greater progress will be made in overcoming them during the next decade.

Problem 1: The Criterion

The first problem is the basic one: we have not known how to assess (measure and evaluate) an operator's performance of meaningful tasks in work situations. The problem of criteria is not unique to performance assessment research, of course, as demonstrated by Smith's (1976) chapter in the Handbook of Industrial and Organizational Psychology. However, the problem at least seems to be more acute in the area of performance assessment, especially where complex operator performances are concerned, and possibly because of the importance associated with the need for good criterion measures as previously discussed.

For example, suppose we were given the responsibility of monitoring a vehicle operator, such as a pilot or an astronaut, in order to specify his current level of performance. What would we measure? How would we proceed? Even if there were essentially no limits on the amount of information that could be acquired and stored regarding (1) the physical state of the vehicle, (2) the physiological state of the operator, and (3) the behavior of the operator, what information could we call "necessary" and on what basis could we use that information for "predicting" future performance? How could we collate the various kinds of information to provide a valid assessment of the operator's current performance, the current level of operator loading, the performance reserves, and the probability of the operator's being able to complete his mission, were additional loads (emergencies) of various sorts to occur? If part of our responsibility included the ordering of a "return to base" that could be completed only an hour or more after the order, how could the information be used to predict the operator's performance during and following that hour?

Ten years ago there was no set of known correct answers to these questions (cf. Alluisi, 1967). The fact is, there is still no set of currently known correct answers. The truth is that we still do not know what we should do, what we should measure, or how we should analyze the data. We still do not know generally how to assess complex human performances in operational systems. We can do reasonably well and, in some cases, even quite well, depending on the system, in measuring system performance. However, even in those cases, unless we know

the transfer functions, we are not able to make valid inferences regarding the operator's performance from our knowledge of system performance. Nor would we be able to make valid inferences or predictions of system performance from knowledge of the operator's performance were we able to measure it, without valid knowledge of the operator-system transfer functions. A decade ago, the situation appeared bleak; today it appears much less so, in part because of the research and development efforts on performance assessment methodology, but for the most part because of the advances in order technologies (especially electronics and computers). For example, it is now quite possible to write the operator-system transfer functions for many complex systems--we do it nearly "routinely" in the case of our more modern and powerful flight simulators. And our capabilities in these regards are still growing. This changes both the probability of making substantial headway on the criterion problem and the nature of the second problem, which has to do with the research strategy or procedures to be followed in directly attacking the problem of criteria for performance assessments.

Problem 2: The Research and Development Strategy

Let us conceive of a continuum that consists of anchor points representing measurement of actual performance on the job at one end of the dimension, through measurement of performance on a simulation of the job, then measurement of performance on a synthesized job consisting of functional elements derived from the job of interest, to the more familiar laboratory approaches that made use of test batteries with well understood measures of performances on specific tasks at the other end of the dimension. The techniques available for our research and development efforts span the entire dimension, and our selection of one or more of the techniques for use in a given program or project may be viewed correctly as a decision regarding research strategy.

Job-performance Techniques. The advantages of basing performance assessments on actual job performance measurements are quite evident. There is no problem with regard to face validity or with the proper population of subjects to be sampled. The disadvantages are equally well known. The subject's behavior during the period of measurement may not be representative of his "typical" performance, the tasks and performances of them may not be representative of the "real" make-up of the job (and, as indicated earlier, without knowledge of the operator-system transfer functions it may be impossible to evaluate the effects of differences in performance on the system operations), and the feasibility of job-performance measurements tends to be relatively low because of considerations such as cost, safety of the subject in the presence of the stresses of measurment, and difficulty in interpreting differences in specific performances as indicative of qualitative differences in over-all performance. It has been, more often than not, the case that these feasibility questions have driven us to the use of expert ratings of performance as criteria, rather than actual job-performance measurements.

But the times and needs are changing and, as indicated earlier, there is reason to press now for reduced emphasis on ratings and for increased emphasis on actual measurements. Fortunately, we probably have all of the necessary technology available. Among the ways we could proceed are the following:

1. We could employ modern job functional analysis techniques to identify and classify different functional aspects of the job.

2. We could employ unobtrusive or job/task embedded measures of performance.

3. We could identify and measure performances on aspects of the job or task that are related to over-all performance in terms of content and predictive validity, as well as in terms of construct validity as the research findings accumulate and converge from other domains such as simulation and mathematical modeling.

4. We could even begin "sensitivity" studies of the effects on performance of various individual differences, using this information to provide guidance in the design or redesign of jobs in order to increase the population of potential workers where that is a necessary, desirable, or cost-effectiveness goal.

In short, we must have made very great progress in performance assessment technology during the last decade, for today the things that we can do are so much greater than they were then! For the first time in my professional lifespan, I do see the possibility of substantial progress in the development of job-performance techniques, performance criteria, and performance assessment technology, based heavily on actual performances rather than dependent primarily (as was the case in the past) on ratings of performance or indirect test measurements.

Simulation Techniques. The advantages of full-scale mission simulation include high face validity and the involvement of the operator in situations that closely resemble the operational conditions to which we want generalization, while permitting essentially unlimited capability for measurements of different aspects of performance. Given the capabilities of the modern simulator, we should be able to use it, and applications of mathematical tools such as factor analysis, to identify components of performance and their contributions to over-all performance. That is to say, we should be able to design studies with simulators that will guide us in the identification of measurements to make in the previously discussed job-performance measurement situations, and thereby begin to make real progress in the development of empirical validity for our performance criteria. Multivariate analytical techniques such as canonical correlations, applied to the measurements of performances in flight simulators and even initially the ratings of performances in aircraft piloting, for example, might even begin to show us the extent to which the performance of tasks like on-pylon turns are or are not related to general piloting skills. Needless to say, the further development of performance assessment technology along lines such as these will impact training, job design, and selection (as called for by Uhlaner, 1972).

There are, unfortunately, some disadvantages associated with the employment of simulation techniques. First, the simulators and their use are expensive, especially in the context of research and development on performance assessment; however, since the potential impact of the findings from such research is so great in cost-benefits terms, there should be little difficulty in demonstrating the cost-effectiveness of this kind of research and development. I would predict that such a demonstration would lead to decisions to favor it over alternative programs with lower potential cost-effectiveness impact.

The second disadvantage is that, as we start developing the design of a simulator for a given task or job, we are faced with the difficulty of specifying the measurement capabilities required to assess the operator's performances in the simulated system. If we measure everything, as might seem logical at the beginning of such a program, we shall be faced with an overwhelming amount of data that is likely to tax the capabilities of even our largest computers. If we

measure less than everything, given our lack of established empirical validity, we risk omitting what might have been the more valid performance. Since there is no real solution to this dilemma, we shall probably take the most practical approach possible and measure as much as is feasible, using expertise, job functional analyses, past practices, and any other basis to identify the most promising of the measurements to take. The typical course followed by research efforts of this nature is scientifically adaptive; that is to say, it is like a self-correcting system in that we start with some relatively large set of measures and, on the basis of early results, discard some of those and then add other untried measures at the next iteration. Hull did not start with all the measures of learning in his first statement of theory.

A third disadvantage of simulation techniques is of greater of lesser importance depending on the objectives of the research and development program; namely, the more faithful the simulation, the better the application of results to the specific operational system simulated, but the worse the generalization to other systems. That is to say, to the extent that the results of our simulation include variances based on specific factors, we shall be able to explain operational systems which also include those specifics, but not other systems. The broader the desired generalization, the more important is this disadvantage.

Synthetic-work Techniques. In order to make inferences regarding the large "g-factor" in work performance and to avoid the last-mentioned disadvantage of simulation techniques, some researchers, myself included (cf. Morgan and Alluisi, 1972), have employed synthetic-work techniques for assessing the effects of various stresses and working conditions on complex human performance. These techniques involve the creation of a job the performance of which can be measured and evaluated; the separate elements or tasks are combined in order to provide for time-shared multiple-task performances of variable operator loadings under controlled laboratory conditions. In terms of advantages and disadvantages, these techniques lie between those of simulation (discussed above) and specific-test techniques (to be next discussed). The test or work batteries employed tend to have relatively high face validity in terms of both content and of acceptance by operational personnel. Because of this acceptance, there is reason to believe that the operator views the test situation as being essentially like the operational and that his behavior tends to be quite similar in the two. Of course, to the extent that the synthetic-work techniques are successful in measuring only the "g-factor" in work performance, generalization of the findings apply to the general, but not to the specific factors in any given operational system.

These techniques still have much to contribute to the further development of performance assessment technology, especially in the area of their specific strength; namely, in assessing the general, or "g-factor" work performances. New synthetic work needs to be designed, developed, and tested, however, because the types of synthetic-work batteries that have been employed up to the present have represented man-machine system types of operations. They have been suitable for use with independent variables of certain classes (temporal, organismic, and situational) applicable to operators in man-machine systems. Different synthetic-work situations appear to be necessary to permit study of the effects of other classes of independent variables such as institutional incentives and disincentives, motivation, the worker's personal weighting of the various aspects of his job, his sense of personal commitment to the work, etc.

Specific-test Techniques. Test batteries such as those employed for selection have typically been designed to measure abilities rather than performance. Still, if one can assume that performance will reflect momentary ability, then one can conclude that specific-test techniques could be employed to provide measures of performance--especially to measure the changes in performance that could be expected to result from the application of independent variables such as stress, work load, or drug-induced state changes in the worker.

The use of a test battery consisting of a number of appropriately selected or designed individual tasks has some very clear advantages over the other techniques discussed. First, performance on each individual task can be assessed rather exactly. Secondly, these performances theoretically can be generalized to other situations in which the abilities measured by these tasks are required for task performance. Thus, if we are able to describe the operational situation of interest in terms of the individual tasks (or factors) included in the test battery, we should be able to generalize our test results without too much difficulty. In other words, the generality of test-battery performance to operational-system human performance depends only on (1) the availability of a taxonomy of the tasks that go to make up the performance tasks in the operational system, (2) a task analysis of specific jobs of interest in the operational system in terms of this taxonomy, and (3) appropriate weightings of the representative tasks in the test battery according to the task analysis. This promise of generality based on a kind of "chemistry of tasks" must be considered one of the principal advantages of this approach and, of course, it was the principal stimulus for the work on task taxonomies cited earlier as being considered so very important for advancing performance assessment technology a decade ago (Fleishman, 1975).

Aside from the fact that we have made very little headway on the task-taxonomy problem, specific-test techniques have somewhat formidable disadvantages. First, they are generally the poorest of the techniques in terms of face validity, especially from the viewpoint of the subject, operator, worker, or ultimate user of the research and development findings. Second, and somewhat related to the first, there is little or no resemblance between the test situation and the operational, and this raises serious concerns regarding the nature of the behavior observed. If the operator or subject approaches the test situation differently than he does the operational, it is not only possible but also highly likely that his behavior will be affected. If he is more highly motivated in the one than in the other, the results obtained in the test situation may not generalize properly to the operational. If he takes the one situation seriously, but responds to the other as to a parlor game, we may not be able to generalize at all properly from the one situation to the other. Finally, and rather fundamentally, "test behavior" attitudes are probably appropriate to, or at least not detrimental to, the measurement of abilities or capacities where maximum short-term output may be expected, but they are probably less appropriate or actually detrimental to the measurement of performances or "work behaviors" that are influenced by variables other than ability--e.g., by pacing for continuation over days, weeks, months, or years; retention of performance reserves generally, but willingness to expend these reserves under "emergency" conditions; etc.

Even so, these specific-test techniques represent our most soundly-based, quantitative research tools for the study of performance assessments, and they have proven quite successful in prior applications to selection, classification,

training, and assignment--or training and personnel technology--requirements. It would probably be quite beneficial to tie in the use of these techniques with further developments in the uses of the others cited, especially with the job-performance and simulation techniques. In fact, by so doing we might even make substantial progress on the problem of task taxonomies as a "spin-off" phenomenon rather than as a core problem for research.

FUTURE ISSUES FOR PERFORMANCE ASSESSMENT TECHNOLOGY

As pointed out by the Defense Science Board Task Force on Training Technology (Alluisi, 1976a, 1976b), the Services have pioneered in (1) the use of complex simulators to train personnel to operate and maintain major weapon systems, (2) self-paced personalized methods of instruction, often computer assisted or managed, (3) performance-oriented training, and (4) managing the training of very large numbers of individuals. Without question, they are destined to pioneer further in the development of performance assessment technology. The advances that can be made will benefit not only the Services, but also the civilian community to an extent at least equal to that of the development of group testing in World War I and its subsequent stimulus for training and personnel technology. Some of the issues important for the development of performance assessment technology during the next decade are already quite apparent; among them are the five discussed in the remaining five subsections below.

Prima Facie Discriminatory Personnel Practices

Given the "Law of the Land," as represented by the Civil Rights Act of 1964 and the U. S. Supreme Court rulings in the Griggs vs. Duke Power and the Albemarle vs. Moody cases, and given the commitment of the Department of Defense and Military Departments to uphold and adhere to the "Law of the Land," one can predict some major changes in past practices under the MOS approach to the military manpower personnel management systems of the Services. These practices constitute, in the judgment of some, prima facie cases of discriminatory employment practices. This can be illustrated by an example based on the employment and utilization of women by the Services. Note, however, that the case made for women is equally applicable to any of the minority groups covered by the law, and that our use of this one group in our example is based simply on our judgment that the illustration is clearer with that group. Also, for purposes of the illustration, let us assume that, on some relevant physical dimensions (height, weight, strength, reach, etc.), human males constitute a normal distribution, 95 percent of which is above the median of the normal distribution of human females. For purposes of simplicity, let us assume that the standard deviations of the two distributions are equal (an assumption that is probably not valid). Given these assumptions, Table 1 shows the percentage of males and the percentage of females above select-points on the hypothesized scale of physical characteristics. For example, the table shows that if a given physical strength is met or exceeded by 95 percent of the males, it it met or exceeded by (only!) 50 percent of the females.

Were we to say that the job requirements were such that the physical (e.g., strength) requirements could be met by only 50 percent of the males, and if we were even able to establish this as a valid requirement, we would not have a legal basis for excluding women from the job since 5 percent of the females could meet that performance criterion. Yet, in the past, our MOS-classification system has been based on just such exclusion principles.

Table 1

PERCENTAGE OF MEN, WOMEN, AND TOTAL POPULATION WITH PHYSICAL CHARACTERISTICS SUCH AS STRENGTH ABOVE VARIOUS HYPOTHETICAL CRITERION LEVELS

Males	Females	Both
95%	50.0%	72.5%
90	35.8	62.9
80	21.0	50.5
70	13.1	41.6
60	8.2	34.1
50	5.0	27.5
40	2.9	21.5
30	1.5	15.8
20	0.7	10.3
10	0.2	5.1
5	0.1	2.5

Note, too, that it is not sufficient to set the criterion for selection at the point mentioned (or any other point, for that matter) on the basis of tradition, past practice, or even expert judgment. It must be demonstrated in a compelling manner by properly conducted performance assessment studies that this is a valid requirement for successful job performance. Thus, even if the hypothetical MOS were open to both genders, if the criterion employed for selection were such as to admit 50 percent of the males and only 5 percent of the females, it would constitute a prima facie case of discriminatory employment practice, and the burden of proof regarding the validity of the selection criterion would fall on the Service in question, not on the individual member of the group who has charged the Service of discrimination.

The next question, which, to my knowledge is not covered by current law, is easy to predict. It has to do with the need for the job to be designed the way it is --i.e., can the job be modified without loss in general system performance so as to be less factually discriminatory? Can the real requirements for the job represented by the MOS be changed to permit more women to qualify for it? This question is not merely a civil libertarian's or woman-liberationist's pipe dream. It has real significance in defining the size of the population or "manpower" pool available to the Services for this and equivalent MOSs. Thus, as the job is redesigned to include additional percentages of the male population in the "manpower" pool, it will also include percentages of the female population in the "womanpower" pool, with these latter rising at a faster rate of increase than the former. As the third column of Table 1 shows, when the criterion is set at the point that 20 percent of the males can meet it, 10.3 percent of the total (male and female) population constitutes the potential employment pool. If the criterion is moved so that double the percentage of males qualify, more than twice as many persons in the general population, or 21.5 percent, now qualify.

If it is moved so that 80 percent of the males qualify (4 times the initial rate of 20 percent), 50.5 percent of the total population will now qualify (nearly 5 times the initial 10.3 percent). I believe we will have to deal with questions of this sort during the next decade, probably quite early in the decade.

When our attention is turned to the possibility of redesigning jobs (or MOS requirements), another possibility generally ignored in the past will make itself evident. Namely, the possibility of purposely designing jobs with increased tolerance for individual differences with the expectation that trade-offs can be, and will be, effected in the performances where people work together. After all, most of what has to be accomplished by what we have come to call "work" cannot be accomplished by single individuals working alone. Rather, humans perform their work as members of crews, groups, teams, and units (CGTUs). Why must each job in a CGTU be designed as though there were to be no individual variation in its performance? We know that to deny individual differences, even in skilled performances, is to deny reality. So why cannot we reconceptualize "jobs" not as fixed, individual task combinations, but rather as flexible combinations that recognize the reality of the trade-offs that actually do take place. We do it nearly instinctively in our most basic team unit, the family. The divisions of labor among the members of different families differ not only according to the capabilities and skills of the members, but also according to their desires and motivational levels. Nor are the divisions constant over time, so that one member of the family may perform a function that is usually assigned to another member at a given time. The same sort of trade-off occurs in working CGTUs, and we have demonstrated it in the laboratory with group-performance task measurements of subjects who were working during the course of illness with infectious diseases. I believe this area of CGTU-job constituents to be the most challenging of this general issue, but I do not believe we are likely to make great progress before the end of the next decade. The prior goals to be reached, as discussed just previously, are too demanding in both priority (immediate necessity to make progress) and difficulty (time and resources necessary to achieve reasonable objectives) to permit any great deal of effort to be devoted to this newer concept. However, my view might be too pessimistic, since closely related progress is likely to be made in dealing with the second issue that will impact performance assessment technology.

Assessments of Individual versus CGTU Performance

Most of the jobs in the world require interactions with other persons and with their task-dependent performances. This is merely to say, in another way, that much, if not most or all, human performance occurs in the context of CGTU performance. One of the major problems that performance assessment technology will face, once it has made its predictable major breakthrough on the performance measurement side, will be that of relating the performances of the individual members of a CGTU to the performance of the group. Anecdotal and testimonial evidence in abundant supply seems to tell us that the two kinds of performance are not equated, nor even in all circumstances highly correlated. As we increase our capability for assessing individual human performance, the need to assess the contribution of this performance to the over-all CGTU's performance will become ever more pressing. We know this now, for it is merely to say that a military unit's "combat readiness" is not assured with the assignment of "MOS-qualified" personnel in all positions. It takes something more to provide the coordination, cooperation, and cohesion that is characterized as required for

"team performance." The point is that our level of measurement sophistication and capability in the area of CGTU performance is considerably lower than in the area of individual human performance, and nearly nonexistent in the relating of the two.

Relating Human Operator and System Performances

A closely related issue, and one that was alluded to earlier in the discussion of some of the characteristics of performance assessment research, is that of the relation between the human operators' and the system's performances in a given system. As was indicated in the earlier discussion, it will be necessary for us to make explicit the operator-system transfer functions if we are to be able to make inferences about the one from the other. This becomes even more complicated when the system-performance assessments are dependent on not a single operator, but on the several members of a CGTU. I shall dwell no further on this issue. It has been with us for a long time, and real progress is being made, principally through mathematical models and computer simulations.

Fidelity of Simulation

A fourth issue has to do with the employment of simulators in (1) performance assessment research and development, (2) training and training research and development, and (3) skill maintenance and the development of tactical concepts and doctrine. Specifically, it has to do with the necessary or even desirable or optimum degree of fidelity for any given purpose, granted that there is a cost-fidelity trade-off, and that for some purposes, such as early stages of procedural training, a simpler instrument of lower fidelity can be more beneficial than a more complex instrument of higher fidelity. Research will have to be conducted to determine optimum degrees of simulation fidelity for different purposes, as indeed is now underway in the Air Force with the ASUPT flight simulator for certain flight training purposes.

While considering the matter of fidelity of simulation, one should remember that the underlying questions have to do not merely with fidelity per se, but also with the nature of the "fidelity" desired or required. Most often when we think of the fidelity of a simulator, we are likely to conceive of the question in terms of physical fidelity--does the man-machine or operator-system interface faithfully reproduce the system or equipment being simulated? But there is another kind of fidelity that might prove even more important; namely, functional fidelity--does the simulator require the operator part of the operator-system complex to perform the functions required of him? Because functional fidelity is generally easier and less costly to achieve and maintain (especially as changes are made in the system being simulated), there are real benefits to be derived by placing as much emphasis on its use as possible. Of course, the emphasis should not be at the price of reducing to any substantial degree the effectiveness of the simulator for its purpose. This is obviously an issue in the performance assessment arena, and it just might provide spin-off findings applicable to the "criterion" and "task taxonomy" problems cited earlier.

Measurements and Predictions of Operator Loads

The fifth, and final, issue that I shall mention is more closely related to the human factors engineering field than any of the other four (with the possible exception of parts of the first, which had to do with direct development of

performance assessment technology for use in preventing discriminatory personnel practices). This is the development of methods for measurements, specifications, and predictions of operator loads in complex systems. It has to do not only with the measurement of momentary loads during operations, but also with the identification of potentially critical workload levels during system operation, simulation, and testing, and even with the prediction of capabilities based on workload estimates during system conceptualization and design.

The potential cost benefits of success in this issue would be, I believe, quite substantial. Although I do not have the data base necessary for a proper consideration of the cost-effectiveness aspects of the issue, there can be little question of the potential. First, with personnel costs in the Services now more than half of the annual outlay, improvements in personnel utilization can be expected to provide benefits in terms of reduced personnel costs. Second, with weapon system unit costs as high as they now are, the reduction of system failures and equipment losses through avoidance of errors based on operator overload would provide considerable benefits in terms of cost avoidances for replacements.

CONCLUSION

The following remarks provide a very rapid recapitulation:

The trend over the last decade has been towards greater emphasis on performance assessment technology, especially in relation to other methods of making inferences about an operator's or worker's performances of work on-the-job in the field rather than on a test in a laboratory. This trend has been further reinforced by law and court rulings that indicate selection tests (for employment and promotion) must be based on carefully defined and validated job-performance criteria.

The characteristics of performance assessment research and development still require emphasis on the criterion problem. This is due to two reasons. First, strategies are based on various mixes of alternative methodologies which, in turn, are based on job-performance techniques, simulation techniques, synthetic-work techniques, and specific-test techniques. Second, there is promise of substantially greater progress during the next decade than was the case during the past decade, in part because of applicable developments in computers, mathematical modeling, simulation technology, and capabilities for job-performance measurement instrumentation.

Among the major future issues for performance assessment technology are:

1. Its convergence with selection, training, and job design technologies in order to provide for nondiscriminatory personnel practices and to increase the available pool of potential candidates for military (and other) service.

2. Its attending to crew, group, team, and unit performances as well as individual performances.

3. Its relating system performance with the performance of the human operator, principally through development of the operator-system transfer functions.

4. Its attention to questions of the optimum (most cost-effective, desirable, or necessary) degree and kind of fidelity in simulation for different usages.

5. Its application in measuring and predicting operator workloads and in specifying optimum loading levels.

The final lesson drawn from a study (Alluisi, 1976b) of Defense training technology is applicable here: "The success of an R&D area within the DoD depends in the final analysis on the activities of that R&D community." If the performance assessment R&D community can present its case in the proper management language of potential cost-effectiveness benefits, the case will be compelling and it will win support to "do its thing" and thereby benefit the nation.

REFERENCES

Alluisi, E. A. Methodology in the use of synthetic tasks to assess complex performance. Human Factors, 1967, 9, 375-384.

Alluisi, E. A. (Chrm.) Summary report of the Task Force on Training Technology. Washington, DC: Defense Science Board, 1976. (a)

Alluisi, E. A. Lessons from a study of Defense training technology. Journal of Educational Technology Systems, 1976, 5, 57-76. (b)

Alluisi, E. A. and Morgan, B. B., Jr. Engineering psychology and human performance. In Annual Review of Psychology, Vol. 27. Palo Alto, CA: Annual Reviews, Inc., 1976. pp. 305-330.

Chapanis, A. The relevance of laboratory studies to practical situations. Ergonomics, 1967, 10, 557-577.

Chiles, W. D. (Sp. Ed.) Methodology in the assessment of complex performance. Human Factors, 1967, 9, 325-392.

Fleishman, E. A. Toward a taxonomy of human performance. American Psychologist, 1975, 30, 1127-1149.

Morgan, B. B., Jr. and Alluisi, E. A. Synthetic work: Methodology for assessment of human performance. Perceptual and Motor Skills, 1972, 35, 835-845.

Singleton, W. T., Fox, J. G. and Whitfield, D. (Eds.) Measurement of man at work. London: Taylor and Francis, 1971.

Smith, P. C. Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976. pp. 745-775.

Uhlaner, J. E. Human performance effectiveness and the systems measurement bed. Journal of Applied Psychology, 1972, 56, 202-210.

ABOUT THE AUTHOR

Dr. Earl A. Alluisi is currently University Professor of Psychology in the Performance Assessment Laboratory of the Department of Psychology at Old Dominion University in Norfolk, Virginia. He has previously worked in industry, with the Lockheed Missiles and Space Company and the Lockheed-Georgia Company, in research, with the Army Medical Research Laboratory and the Stanford Research Institute, and in the academic community, with Ohio State University, Emory University, and the University of Louisville. Over the years he has served as teacher, researcher, and administrator. He is credited with more than 200 publications and research reports, and was recipient of the Jerome H. Ely Award of the Human Factors Society in 1970, and of the Franklin V. Taylor Award of the Society of Engineering Psychologists in 1971. He is a former president of the Society of Engineering Psychologists (Division 21) of the American Psychological Association, and of the APA's Division of Military Psychology (Division 19). He is a fellow or member of numerous professional organizations, and currently serves on the Publications and Communications Board of the American Psychological Association, and as Chairman of the Faculty Council at Old Dominion University.

PLANNING FOR PERFORMANCE MEASUREMENT R&D: U. S. ARMY

Milton S. Katz
U. S. Army Research Institute for the Behavioral and Social Sciences

ABSTRACT

This paper describes the history, status, and projections of Army performance measurement research and development. A large-scale performance measurement application--Skill Qualification Tests--is taken as a case of concurrent and continuing research development and implementation.

RESEARCH AND DEVELOPMENT IN THE ARMY

The U. S. Army Research Institute (ARI) is a field operating agency of the Deputy Chief of Staff for Personnel (DCSPER) and a developing agency for personnel performance and training research, development, test, and evaluation (RDT&E). It performs the RDT&E to improve operational practices and procedures in the areas of personnel and management systems, educational and training systems, and human factors in system development and operation. ARI research is concerned with the whole person functioning in the Army system: (1) the effects of individual variables on adaptability and functioning in a wide variety of Army jobs in different organizational contexts; (2) the effect of work environments, including system complexity and automation, on individual and group performance; and (3) the requirements for, and effects of, training and evaluation at all skill and operational levels.

ARI's research is contained in the Science and Technology portions of the Army RDT&E Program, which include basic research (RDT&E category 6.1), exploratory development (RDT&E category 6.2), and advanced development (RDT&E category 6.3A). Work in 6.1 and 6.2 address identified areas in which there is insufficient scientific knowledge; effort is generally expended in formulation of scientific principles and identification parameters. Work in 6.3A involves application of scientific knowledge gained from 6.1 and 6.2 efforts to current or potential field problems and to demonstration or validation of operational utility. This effort may be expended in response to user requirements as stated in: (1) the Science and Technology Objectives Guide, (2) a Human Research Need Advisory Statement, or (3) a jointly approved Department of Defense (DoD) program.

Annually, a two-volume publication, entitled ARI Science and Technology Program, is issued. Parts I and II, respectively, describe the status and plans for efforts in 6.1 and 6.2 and those in 6.3A. Five-year plans, which extend through Fiscal Year 1981, are currently available for most 6.3A projects. Policies governing ARI's operations and relations with other agencies are detailed in Army Regulation 70-8, dated 28 October 1976.

Projects are conducted and managed by in-house ARI scientists (research psychologists, educational technologists, psychometricians, computer scientists, and other pertinent professional and technical specialists) augmented by selected research and data collection contractors or grantees. Diverse geographic requirements for coordination, implementation, and evaluation of necessary project activities are materially

supported by collocation of ARI field units and liaison personnel with U. S. Army Training and Doctrine Command (TRADOC) headquarters and components (including service schools and training centers), and Army test and operational commands. In addition to programmed research efforts that are coordinated routinely with major commands and field operational components, ARI supplies requested expert Technical Advisory Support (TAS) to address unanticipated or emergency human resource problems.

The Army has long been a leader in developing performance-based instructional technology and applying it to large-scale training programs. The Army has been effecting institutional change to introduce more accurate accountability of training effectiveness in terms of performance on critical job tasks. Other programs--in DoD, civilian industry, and civilian education--are also developing and adopting performance-based technology. The Army works closely with other government organizations and services through cosponsorship and through participation and presentation at meetings, such as the Military Testing Association, the Society for Applied Learning Technology, HEW's Work Conference on Analyzing Materials for Instructional Materials Developers, Interservice Group on the Exchange of TM Technology, and the Military and Army Operations Research Conferences. In addition, there is direct contact between DoD research laboratories and civilian contractors not only through contracts and grants, but also through participation in professional society activities and through government and professional publications of research reports.

Major Research Thrusts in the Past

For most of the past three decades, Army performance measurement research and development has been directed predominantly to technical and practical problems of screening, selection, and assignment. The major efforts have been toward development and refinement of technology and instruments to measure aptitudes, educational achievement, intelligence, and other general and specific skills and characteristics of candidates, trainees, and incumbents to maximize the efficiency of personnel selection, training, assignment, and utilization. The universal draft was in operation, and the force of the approach was to economically screen and develop Army personnel who would be likely to succeed.

Project 100,000, in the late 1960s, lowered the Army's entry standards and admitted a group of recruits whose abilities and performance strained both the training system and the testing system. The immediate reaction of the Army, as well as of other services, was to reengineer training and jobs to enable persons with lower general mental ability to learn and perform military jobs adequately.

Concurrently, instructional systems engineering, a product of industrial, educational, and military development, was gaining impetus and seemed ripe for application in the Army. The Army's 1968 version, called "systems engineering" (Army Regulation 350-100-1), incorporated the principles of:

1. Identifying training needs on the basis of job requirements.
2. Conducting training analyses to identify content, procedures, and standards.
3. Maintaining quality control through evaluation.

The regulation laid out the priorities and procedures for designing or redesigning new or old courses over a period of "as much as 5 years." Needless to say, the ambitious program of redesigning even existing courses had not been completed 5 years later. Nor was much of what had been done a success. Even willing trainers and training managers did not have sufficient skills to implement systems engineering. Incentives to comply were weak or nonexistent.

Further development of systems engineering of training and evaluation resulted in a triservice Instructional System Development Model. The Army's version includes five volumes, totaling about 1,000 pages, but it is difficult to implement because procedures are often not operationally designed, and because the intended implementors are still heavily influenced and hampered by their own training and accustomed ways of dealing with instruction and performance measurement.

Before Project 100,000 and systems engineering, the focus of research and development was classroom instruction directed by an instructor. Research efforts in the Army were aimed toward improving training conducted in classrooms or, at least, reducing its cost. Research on performance measurement, accordingly, was conducted in the context of classroom instruction. Such performance measurement evaluated the effectiveness of instruction, but assumptions about success in job performance were purely speculative inferences with no corroboration in real life.

New Army research initiatives undertook to investigate the usefulness of hands-on training to develop competence in job tasks, and to devise and evaluate hands-on performance tests as measures of proficiency in job tasks. The research effort was sufficiently successful to gain support for a large-scale effort called ATC-PERFORM.

The principles of ATC-PERFORM were not new. In fact, they hearkened back to AR 350-100-1 with more emphasis on job and performance contexts. The principles were to employ:

1. Performance-based instruction.
2. Absolute, go/no-go standards.
3. Functional job context.
4. Individualization.
5. Feedback of results.
6. Quality control.

The effort was limited to entry-level job training conducted in an institutional setting. The primary focus was on performance-based training, but performance testing was not developed systematically. In addition to the institutional changes, which brought researchers into intimate contact with full-scale training situations and involved school personnel in developing and implementing innovations, tangible products of lasting value resulted. SMART (Soldier's Manual Army Testing) books proved to be a mechanism for sustaining performance-based training and testing in the schools, and served as a developmental model for the current Soldier's Manuals. Other products include manuals for trainers on how to conduct performance-based training and testing.

Taking into account imminent budget and resource cuts, discontinuation of the draft, high rates of personnel turbulence within the Army, long-recognized inadequacies and inefficiencies in training, and less than satisfactory combat readiness of troops and units, the Army set out to implement the Enlisted Personnel Management System (EPMS) in 1974. The progressive changeover involves approximately 1600 skill levels in 300-400 Military Occupational Specialties (MOSs) in 39 career management fields. This system requires restructuring and consolidation of MOS classifications to reduce the number of specialties and senior noncommissioned officers. It further presses for revision and decentralization of training and testing systems so that they will ensure:

1. Training pertinent and necessary to the job and the mission.

2. Testing that is a realistic and valid measure of job skills mastered.

3. Optimal training, assignment, and evaluation.

4. Fair and equal treatment of soldiers.

Performance Measurement Research Applied to Operational Problems

The three essential building blocks needed to support a system such as EPMS are:

1. Soldier's Manuals

   Specification of Critical Job Tasks, Conditions, and Standards.

2. Performance-Based, Exportable Training, and Job Aids

   Field and Technical Manuals.

   Correspondence Courses.

   Training Extension Courses.

   Integrated Technical Documentation and Training.

   REALTRAIN.

   MILES.

   Others.

3. Skill Qualification Tests (SQTs)--Army Training and Evaluation Program (ARTEP)

   Criterion-Referenced Performance Measurement.

It is clear that establishment of performance standards and criteria pervade the entire system. Without criteria and a satisfactory means of determining whether or how thay have been met, the system remains an open loop. Training that has no demonstrable impact on job performance or on mission accomplishment fails to meet the operational requirement. Performance measurement that gives no sure sign of job mastery results in an inequitable personnel system. Accordingly, Skill Qualification Tests (SQTs) are the kingpin of the EPMS system. They measure and drive the training and personnel management systems.

Most, if not all, of the performance measurement research issues entailed in specifying critical job tasks, conditions, and standards, and in developing and implementing performance-based training and job aids, whether for individuals or units, are embodied in the development and application of SQTs for individuals, and the ARTEP for units. Consequently, a close examination of problems and solutions in the large-scale application of criterion-referenced individual performance testing will highlight gaps in our research base, and point future directions.

The urgency of the operational demand for Skill Qualification Tests precluded a measured, systematic, long-term research program in advance. Almost nothing generated to date by the research or academic community was adequate to the job of developing or implementing criterion-referenced, performance-based, job-relevant evaluation. The conventional logic, statistics, construction methods and wisdom tended to be irrelevant, counterproductive, and off the mark. The researchers participating in this effort had to discover how to build tests to measure performance, to establish their validity, to measure and ensure their reliability, and to make them feasible to administer and process.

THE CASE IN POINT--SKILL QUALIFICATION TESTS[1]

Overview

Skill Qualification Tests (SQT) have been developed to replace Military Occupational Specialty (MOS) proficiency tests as measures of ability to perform Army enlisted jobs. SQTs are performance-based, criterion-referenced measures of job proficiency, consisting of precisely defined tests of tasks, all of which are critical and necessary to performance of the job. The criterion-referenced approach provides an explicit relationship between job requirements and test content in that job requirements dictate content of SQTs. The SQT development process requires that tests be reviewed by subject matter experts and validated on representative job incumbents to assure that test content is job relevant. Test standards of acceptable levels of performance are also based on job requirements and test content. Performance standards are based on behaviorally derived absolute scoring standards, and not on performance relative to other soldiers who take the test. For these reasons SQTs are justifiably viewed as criterion-referenced tests of job proficiency.

A criterion-referenced testing system offers two significant advantages that are not available in traditional testing programs. One is that test content can be made public in advance of administration. There are no reasons to keep test content secret in a testing program based on explicit linkages between test content and job requirements. Advance knowledge of test content results in an equitable and open system. Everyone has an equal opportunity to acquire proficiency on the specific job tasks known to be included in the test.

---

[1]THE CASE IN POINT--SKILL QUALIFICATION TESTS section was written by Milton H. Maier and Stephen F. Hirshfeld, of the U. S. Army Research Institute for the Behavioral and Social Sciences (Individual Training and Skill Evaluation Technical Area).

The second is that a criterion-referenced approach allows personnel management decisions such as those involving promotion, selection, and advanced schooling to be based on performance standards instead of personnel quotas. In more complicated situations involving the merging or splitting of job specialties at higher skill levels, soldiers from different specialties can be compared on their levels of competence rather than their relative standing in the testing group. Criterion-referenced testing of job proficiency has opened new opportunities for both training and personnel management.

Background

The Army has been using tests to measure job proficiency for over 15 years. These tests, called Military Occupational Specialty (MOS) tests, were designed primarily to help personnel managers make decisions of vital importance to individuals' careers, such as proficiency pay, promotion, and assignments. The MOS tests are traditional achievement tests, consisting of 125 multiple-choice items, each with four alternatives. The test content is related generally to the domain of job performance, but there is no definitive logical correspondence between test items and specific job requirements. Each item is scored pass-fail; the total score is the number of items correct, and the total score is then used to rank persons in each job specialty. Therefore, any referencing of test score to test content is immediately abandoned. Because of content limitations, lack of content-score correspondence, minimal diagnostic utility, and the long delay in providing feedback to the field (as much as 1 year after testing), Army trainers have found MOS tests not particularly useful for determining training requirements, measuring individual and unit performance, or assessing training readiness.

Army training during this same period, especially in the late 1960s and early 1970s, was undergoing a major revolution. Performance-based training and testing, based on critical job tasks and criterion-referenced standards of performance, were being implemented in entry-level training courses. Training objectives were operationally defined by the performance tests given during the course, and the tests were made public to students as well as instructors. The content of these tests was always directly relevant to the job. The tests themselves were used to drive the direction of training.

Tests, because of their function in maintaining accountability, are effective instruments in bringing about institutional change. Test content helps implement doctrine about the way jobs are to be performed, and are helpful in defining training requirements and standards. The public nature of the tests helps focus attention on the critical elements of the job, enables effective use of soldiers' time in preparing for tests, and thus improves individual readiness.

The new criterion-referenced tests, which have evolved, are called Skill Qualification Tests (SQT). Their implementation is profoundly influencing the entire Army community. The new testing procedures are forcing soldiers, training managers, personnel managers, and research support personnel to rethink and often redefine their functions.

Requirements of Skill Qualification Tests

The basic requirement of SQTs is that the tests be job relevant. The test content must be based on critical job requirements, and the test scores must be accurate measures of ability to perform critical job tasks.

The job relevance of SQTs is accomplished by basing them on Soldiers' Manuals. These manuals identify the critical job tasks, the behaviors required to perform the tasks, the job conditions, and the standards for performance. They define the jobs in that they list all the tasks soldiers in a job specialty are responsible for performing. Since SQTs are based on Soldiers' Manuals, the SQTs are job relevant.

SQTs are used by both training and personnel management to help make important decisions affecting the career development of soldiers. Both training and personnel management need timely and accurate information about how well individuals are performing; the former to determine training requirements of individuals, and the latter to help determine whom to promote, reclassify, or reassign. Although training and personnel management have a need for the same kind of information, their immediate requirements are not identical.

Training managers base their immediate training requirements on the specific tasks performed in their units. The job relevance of tests for specific assignments, therefore, is the primary consideration from this point of view and it is defined in terms of the tasks that soldiers perform in their assignments. The set of tasks performed in an assignment is generally a subset of the tasks required in a specialty. The task is a convenient unit for determining training requirements because tasks are observable, have initiating and terminating cues and have standards of performance that can be reasonably well specified. Decisions about proficiency can be made at the task level, and training managers can identify the specific tasks on which soldiers need training. If the tests measure performance on the specific tasks for which the training managers have responsibility, they are serving their basic purpose.

Personnel managers are also concerned with the job performance of individual soldiers, not only in specific assignments, but in all the tasks in a specialty. For example, performance in a specialty such as Infantryman cannot be inferred from the limited set of tasks that constitute the specific job assignment of rifleman or of radio-telephone operator. Personnel managers, therefore, have a need for information based on a standard set of tasks for each specialty. All soldiers in a specialty need to be evaluated on the same set of tasks to enable fair decisions about which soldiers to promote, retain, or reclassify.

The need for a standard set of tasks in each specialty imposes additional testing requirements for feasibility and acceptability. The test scores should not be affected by when or where the test is taken, nor by whom it is administered and scored. The testing conditions, as well as performance standards, should be standardized.

The requirement for Army-wide standardization at the present state of the art in testing means that initially most of the test content is in the paper-and-pencil mode rather than hands-on performance tests. Paper-and-pencil tests generally lack the apparent job relevance of hands-on performance tests, and therefore an additional requirement is imposed to assure that the tests are acceptable to examinees, supervisors, and commanders as valid measures of job proficiency.

Job relevance of the tests is the basic requirement for both training and personnel management, even though the definition of job relevance may have somewhat different meanings for the two purposes. For training purposes, the focus is on the subset of tasks performed in the specific job assignment whereas, for personnel purposes, the interest is on the entire set of tasks in the specialty.

The SQTs are designed to serve the requirements of training and personnel management. Because of their divergent immediate needs, critical issues arise in how SQTs are developed, scored, and used. These issues constitute an urgent, high-density focus in current and future Army research and development in performance measurement.

Development of Skill Qualification Tests

The Skill Qualification Testing (SQT) program is a large-scale effort to provide valid and efficient measures of job proficiency. Because of the strategic importance of Skill Qualification Tests to both training and personnel management, high-level policy decisions were made about test content, validation, and scoring. The general requirements of the program are that tests must (1) be fair and feasible, and (2) have validity demonstrated in advance of operational use.

Fairness and Feasibility of the Tests. Fairness means that all soldiers have an equal opportunity to demonstrate their true level of job competence. Test content must be based on actual job requirements, and testing conditions must be sufficiently constant throughout the Army so that scores obtained from administration under varied conditions are not noticeably different. Tests given in Alaska, Panama, Germany, or the contiguous states must all be administered under similar conditions. In addition, all persons administering and scoring the tests must be able to do so accurately and objectively. Still another requirement is that the tests be acceptable to soldiers and knowledgeable experts as fair measures of ability to perform critical job tasks. Therefore, fairness attends to requirements of both training and personnel management.

Feasibility requires that the tests be suitable for administration in all types of units and environments. Equipment, terrain, personnel, and all testing material must be readily available. Another aspect of feasibility is that testing time must be reasonable, with up to 1 day allowed for testing each soldier.

The requirements that SQTs be fair and feasible put severe limitations on the use of hands-on performance tests. The history of performance measurement is that scoring accuracy and standardization are difficult to attain. One resolution of the fairness and feasibility requirements is to have several kinds of testing. Under present policy, all SQTs contain a written component, and some contain a hands-on component. Four hours of testing are allowed for the written component, and up to 4 hours for the hands-on portion. A third component, called performance certification, which is essentially an observational evaluation of actual job performance, may also be included.

Therefore, an SQT may include up to three distinct types of tests, each with its own inherent strengths and weaknesses. A combination of these tests is an operational answer to the fairness and feasibility requirement.

Hands-on performance tests are most desirable. They are a form of structured observation in which a scorer evaluates an individual on a set of performance measures (observable behaviors). Advantages of hands-on testing are obvious: it tests actual performance, has high fidelity to the job, allows for immediate feedback, and has high face validity to examinees. However, considerable developmental effort is required to ensure scoring realiability and standardization of conditions. It also is expensive in terms of equipment, personnel, and time; in other words, feasibility is often a problem. In order to ensure feasibility,

there is a natural tendency to truncate tests of tasks by shrinking the boundaries. Unfortunately, this may be at the expense of the validity of the test. For these reasons it is extremely difficult, if not impractical, to initiate a large-scale hands-on testing system for an organization as large as the Army. Therefore, a hands-on component constitutes a subset of an SQT.

The performance certification component, an alternative form of hands-on testing, covers tasks that are too long, complex, or resource-intensive to include in the hands-on component, and do not lend themselves to testing in a written mode. Performance certification tests are to be administered and scored by soldiers' supervisors in the normal job setting. Although performance certification allows greater flexibility and avoids some of the feasibility problems encountered in a hands-on component, there are problems in ensuring reasonable standardization of job testing conditions across individuals and standardization of scoring by supervisors. Sound methods for addressing these problems are needed to make performance certification a significant and powerful portion of an SQT.

The decision to include a written component imposes careful consideration and analysis of what criterion-referenced measurement means in this context. Since the focus of SQTs is on ability to perform critical job tasks, that aspect must be retained. Each written test of a task is to consist of a set of items, where each item is designed to measure an essential behavior or step in performing the task. For tasks that require primarily "mental" or writing skills, as in the supply and administration fields, written tests of tasks are often similar to or indentical with the behaviors required on the job. Then the standards for ability to perform the test of the task can be reasonably close to those on the job. For tasks requiring psychomotor skills, written test items only simulate actual job behaviors, and the setting of realistic standards indicating ability to perform the task entails a more remote inference. To help approximate realistic job conditions, written items may have multiple correct responses and variable number of alternatives. This added flexibility, however, increases the difficulty of developing appropriate methods for setting standards. The determination of reasonable standards for written tests of tasks is one of the most difficult issues in the SQT program.

Because Army jobs and training programs are structured in terms of critical tasks, the appropriate level of analysis for the SQT should also be based on tasks. The concept of "scorable unit" was invented to help assure criterion-referenced measurement of task performance. A scorable unit is designed to measure ability to perform a specific task or, in the case of complex tasks, a well defined subtask.

Each written scorable unit consists of a set of items, each of which is designed to measure an essential behavior or step in performing the task. Each item is scored pass-fail, and a prescribed number of items must be passed to attain GO on the written scorable unit. A GO is counted as ability to perform the task. Currently, standards for written scorable units require that an a priori number of items be passed. For example, if a scorable unit contains five items, then four must be passed to obtain a GO.

Hands-on and performance certification scorable units consist of a set of performance measures. Each performance measure is scored pass-fail, and a prescribed number of performance measures must be passed to achieve GO on the scorable unit. A GO on the scorable unit is interpreted as ability to perform the task. The standards for GO generally are comparable to what is required on the job.

The requirement that all scorable units be acceptable as fair measures of ability to perform tasks is applied to both the hands-on and written tests. Juries of experts must agree that the written items and hands-on performance measures reflect ability to perform the tasks. Perhaps a safer statement would be that failure to pass the items indicates that the person is not able to perform the task.

Establishing a Correspondence Between Test Content and Job Tasks. The most critical requirement of SQTs is their job relevance. Test content of all SQTs is a sample of critical tasks from the domain of job tasks in the specialty. In this way, the tests have a specifiable and explicit link to the job. For each Army job, there exists a Soldier's Manual that lists the tasks for which a soldier in that specialty is responsible. Therefore, this set of tasks operationally defines the job. Tests to measure performance on specific job tasks listed in the Soldier's Manual are developed from appropriate task analyses. The tests for each task, therefore, are operational definitions of performance on the task. Performance on individual tasks is summed to obtain a total score, which in turn operationally defines job competence. Modern instructional technology, with its emphasis on specification of objectives and verification that those objectives are attained, supports the process for establishing the content and focus of SQTs.

Although the task is the basic level of analysis, the validity of task proficiency measurement depends on the adequacy of the test of the task. By means of detailed task analyses, the set of performance measures or behaviors required for successful performance of the task are identified. These lists of performance measures are all available in the Soldier's Manual. Each item developed to test for task proficiency must occupy a clearly specified relationship to a performance measure required in task performance. Assuming that the set of items developed for a test of a task has been selected in accordance with the procedures described above, one may assume with reasonably high confidence that successful performance of each tested behavior is a necessary condition for successful performance of the task.

How to score the set of items in a written scorable unit to obtain estimates of ability to perform tasks is a complex question. Measurement error is always a problem that must be allowed for. Whether being scored GO on a test of a task requires passing all items included in the test of the task, or some number less than perfection, depends on the nature of the task, the fidelity with which the task can be tested in a written mode, the complexity of the format (e.g., multiple correct responses), and the number of items within the cluster. Use of subject matter experts in reaching such a determination is mandatory.

In the case of a hands-on test of a task, measurement error arising from the use of words is minimized. However, other measurement problems arise. One is that a full performance test of a task generally is not feasible. It may be too costly in terms of time, equipment, and personnel. Therefore, a truncated test of the task is often developed by eliminating some of the performance measures or steps required for the full performance test. By truncating the test, though, it is possible that the tested portion is necessary, but not sufficient, for successful task performance.

Validating Tests Prior to Administration. A first question to be answered was how to define validity. The starting point was the usual definition of validity; that is, that the tests measure what they are intended to measure. In the case of Skill Qualification Tests, the intent is to measure ability to perform

critical job tasks. The content of the tests, therefore, becomes the crucial factor in establishing validity, and must be thoroughly reviewed by experts to ensure that the right behaviors and decisions are assembled in each scorable unit. The first requirement, then, is consistent agreement among experts that the content of the test is based on ability to perform critical job tasks. A second requirement is that the scorable units discriminate between successful performers (masters) and nonperformers (nonmasters). A third requirement applies only to written scorable units: All items in a written scorable unit must be consistent estimators of mastery on the task covered by the entire scorable unit. Thus, the concept of validity focuses on consistency: consistency among expert reviews, and consistency in identifying mastery.

SQTs are constructed and validated by Army agencies that have resident expertise in the job specialties. Generally these are the Army schools, but they also include other agencies, such as the Health Services Command. Since the test content must reflect job tasks, the test developers must have detailed analyses that identify the behaviors essential to successful performance of the tasks. SQTs are developed in the following conceptual sequence:

1. Identify tasks for testing.

2. Identify behaviors or steps essential for performing each task.

3. Develop scorable units to cover essential behaviors of the task, and review scorable units for content validity.

4. Try out scorable units on soldiers to verify accuracy of measurement.

After each step in the process, the products are submitted to higher headquarters for review and approval. The content of the scorable units is fixed after step 3. Scorable units found to be unsatisfactory through tryout on soldiers can be revised, but the content cannot be changed. Test content is fixed through agreement among experts that the content of the scorable units validly measures ability to perform the tasks. The tryout serves only to establish the measurement properties of the scorable units.

The tryout with soldiers is different for the hands-on and written components. For the hands-on tests, the primary concern is to establish that the performance measures can be scored accurately. Acceptable agreement among the scores is considered to be attained when 80 percent of all pairs of rater scores are the same for the performance measures in a scorable unit. If less than 80 percent agreement is obtained, then the performance measures are revised until an adequate level of scoring consistency is attained.

For written tests, the tryout is concerned with establishing the effectiveness of scorable units in distinguishing between performers and nonperformers, and with assuring that all elements in a scorable unit are consistent in estimating ability to perform the task. This tryout helps assure that all items of a scorable unit contribute to measuring each performance.

A final evaluation of the written scorable units is conducted after operational administration of the tests. A representative sample of answer sheets is selected for analysis, and item correlations within each scorable unit are obtained. Items with a positive intercorrelation pattern are retained, and items negatively correlated

are deleted prior to final scoring. When all steps of the review and analysis procedure for the written scorable units are accomplished, their validity as fair measures of ability to perform job tasks is considered to be reasonably well established.

Assumptions in Scoring SQTs

Three sets of assumptions have been made in scoring SQTs to justify using SQTs to help (1) determine training requirements, (2) select soldiers in both single and merged specialties.

Determining Training Requirements. The following, which are required for using SQTs to help determine training requirements, are straightforward:

1. Tasks can be defined--task elements or behaviors can be specified, conditions given, and standards of adequate performance established.

2. Tasks can be measured validly--performance on the task is measured by scorable units, which contain time or performance measures related to task elements, and the sum of the elements passed in a scorable unit indicates quality of performance on the task.

3. Task elements are weighted equally--items or performance measures corresponding to task elements or behaviors are scored as pass-fail, or as one-zero.

These three assumptions serve to provide operational definitions of performance on the tasks measured in SQTs. Although task elements do not have to be weighted equally, research evidence indicates that differential weighting generally does not improve the quality of measurement. A common practice is to give an element greater weight by preparing several items or performance measures for it.

The assumptions needed to help determine training requirements pertain only to tasks taken one at a time. Since the current training philosophy is to train on discrete tasks, no assumption about the interrelationships among the tasks is required.

Selecting Soldiers in a Single Specialty. Using SQTs to help select soldiers in a single specialty does require additional assumptions about the interrelationships among job tasks and scorable units that measure task performance. The same three assumptions about measuring task performance are required; that is, tasks can be defined and measured validly, and task elements are weighted equally.

In addition, three more assumptions are required:

1. Scorable units are weighted equally--all are scored as GO/NO-GO or as one-zero.

2. Test score is the number of scorable units performed correctly--the total score is obtained by adding up the number of scorable units passed.

3. The percent of scorable units passed indicates level of job performance--the percent of scorable units passed corresponds to the proportion of job tasks a soldier can perform.

Given these assumptions, SQTs define the criterion of job proficiency, and the percent of scorable units correct (called percent correct) is a direct reflection of job proficiency. Standards of job proficiency can then be set in terms of percent-correct scores.

Selecting Soldiers in Merged Specialties. In the case of merged specialties, an additional assumption is required about the relationships among the jobs or groups of soldiers. The first six assumptions made in the case of the single specialty result in criterion-referenced measurement for each of the jobs being merged. However, in order to maintain criterion-referenced standards for merged specialties, the additional assumption is required: that the jobs being merged are equal--that is, equal levels of proficiency in the individual jobs are equal to each other in an absolute sense. Stated operationally, all scorable units from all the relevant SQTs are weighted equally. Thus, a soldier qualified in 45N (Tank Turret Mechanic), for example, is equal to the qualified soldier in 45P (Sheridan Turret Mechanic), regardless of the percentage of soldiers in each qualified group. An implication of this assumption that the jobs being merged are equal is that if one qualified group contained 5 percent of a first MOS population while a second qualified group contained 50 percent of a second MOS population, the merged qualified group would contain proportionately more soldiers from the second group.

In the above example, each MOS would be represented in the merged qualified group in accordance with the number of soldiers from each MOS who attained qualifying scores. One MOS may be proportionately overrepresented, while the second MOS is minimally represented or possibly not represented at all. How to use and maintain performance standards for merging MOS is a policy decision, and not a technical question. However, the criterion-referenced properties of SQTs enable rational policy decisions.

An alternative assumption in the case of merged specialties is that the groups, and not the MOS, are equal--that is, equal percentile-rank scores indicate equal levels of job proficiency. The use of percentile-rank scores, which indicate relative standing in a group, facilitates proportional representation of each MOS in the merged qualified group. For example, a policy decision could be made that 40 percent of each MOS be considered eligible for promotion. Such a policy decision might be made if policy makers judged that the jobs were unequal, or that the SQTs were not equally valid criterion-referenced measures of all the merged MOSs, or that the need for proportional representation of the MOS in the qualified group outweighed the need to maintain performance standards. However, if SQTs are scored by percentile-rank, and qualifications are based on percentile-rank scores, then the job performance standards would be given little or no consideration in determining the qualified group.

Benefits from Using Criterion-Referenced SQTs

The change in focus from norm-referenced MOS proficiency tests to criterion-referenced SQTs has enabled training and personnel management to obtain more comprehensive and meaningful information than before. Two major benefits that have resulted from the adoption of the criterion-referenced approach are (1) the public nature of test content, and (2) job performance standards versus personnel quotas.

Public Nature of Test Content. An effective job proficiency testing program should be part of a larger system that includes job requirements and training programs. Modern instructional technology emphasizes the systems approach to training, and a job proficiency testing program is an integral component of the Army's modern training system.

Job requirements are defined by Soldier's Manuals, which list all the tasks a soldier in an MOS skill level is responsible for performing. Soldier's Manuals are distributed throughout the Army for use by individual soldiers and for developing training programs, both resident courses and decentralized training conducted in units. Soldier's Manuals are also used to develop SQTs. Every task tested is in the Soldier's Manual. Once the system becomes fully operational, all components of the Army can know what each soldier should be able to do, is able to do, and should be trained to do. There will be no surprise requirements.

In addition, the SQT Notice gives soldiers advance detailed information about the job tasks on which they will be tested. The Notice lists the specific tasks included in an SQT, tells how the tasks will be tested (written or hands-on), provides standards, and describes the actual test content. All soldiers in an MOS are given equal information about what they will be tested on, allowing them equal opportunity to prepare for the test. Test content, at least in general terms, is public knowledge.

The public nature of test content reduces the need for representative sampling of tasks. One reason representative sampling of tasks is important in the typical testing program is that all examinees are given an equal opportunity to demonstrate their competence. With the SQT Notice, test content can be focused in special areas, such as areas that have high training needs or that are related to new equipment in the field.

The public nature of SQT content also helps establish an integrated training and testing program based on critical job requirements. By selecting test content that focuses on critical job requirements, training efforts will tend to be directed toward these same requirements. Thus, an integrated training and testing system is being developed to meet job requirements.

As long as individuals are tested on the specific requirements of their jobs, there is no advantage to keeping the test content secret. In fact, if the test is directly related to performance on the job, then the proficient individual should already know the test content without the benefit of the information contained in a test notice.

A problem that arises in the typical testing program, where test content is kept secret, is that some individuals have special advantages over others. One possible advantage is that because of favorable job assignments, job tasks and test content are very closely related for some individuals. In the past, soldiers who were working outside of their MOS were at a distinct disadvantage on the test content based on MOS-specific job tasks. The effects of such assignments are minimized in the SQT program because all MOS soldiers are told specifically what content will be included in the test. The prior knowledge about test content tends to equalize opportunities.

In the past, some soldiers have had advantages because they were more familiar with the voluminous references given for MOS tests. Some soldiers did not have

the references available to them, and those who did had difficulty in identifying the critical information within the mass of paper and words. In the Soldier's Manual and SQT Notices, the critical information is distilled and made available to all MOS soldiers. Thus, soldiers with high verbal fluency or with access to specialized information no longer retain such a distinct advantage. Since the critical information is made available to all soldiers in a form readily understood, the opportunities to acquire competence are equally available to all.

Some individuals seem to have a knack for doing well on tests, while others appear to freeze when confronted with a testing situation. Test wisdom is frequently cited as an explanation of why some do better than expected, and test anxiety is given as a reason why some do more poorly than expected. Both of these factors--test wisdom and test anxiety--are undesirable influences because they distort the meaning of test scores. In the SQT program, where everyone has an opportunity to practice for the test, the effects of test wisdom and test anxiety are minimized, and the scores are more likely to reflect true levels of competence.

A factor related to test wisdom and test anxiety is the threat that many soldiers experience when taking tests. The threat may be viewed as having both objective and subjective components. A major source of objective threat arises from the fact that SQTs are used to help make personnel decisions that affect careers. Soldiers who do poorly on SQTs are likely to be penalized, while those who do well are rewarded. The test then, understandably, poses a threat to many soldiers, especially those who are marginal performers or who are not familiar with testing, or who have had negative experiences in school situations.

Subjective components of threat may arise from a variety of circumstances, such as personal characteristics, prior experience with tests, or from a fear of being evaluated. The fear of being evaluated may arise because the rules or basis for the evaluation are not explicit. If soldiers have foreknowledge about the tasks they will be evaluated on and how the evaluation will be conducted, then the subjective threat may often be reduced. Prior knowledge about test content may equalize opportunities for soldiers to demonstrate their true level of job competence by reducing distortion of test scores arising from subjective threat.

The public nature also has the general effect of increasing the validity of the tests. By giving all MOS soldiers an equal opportunity to prepare for the tests, the test scores are more likely to reflect true levels of competence.

Job Performance Standards vs. Personnel Quotas

A criterion-referenced job proficiency test consisting of task-based tests can be scored in terms of percent of tests correct, which is a direct indicator of the percentage of job tasks a soldier can perform and, therefore, a direct measure of level of job competence. The percent of task-based tests correct can be interpreted because standards are specified. The distribution of scores is not a relevant consideration in interpreting the meaning of the scores.

For each task in an SQT, two categories of performance are established--qualified and not qualified. Therefore, SQTs provide GO/NO-GO decisions on task performance. Soldiers either meet these standards or they do not. The total SQT score is the sum of all scorable units passed, which provides continuous scores ranging from all scorable units correct to none, or 100 percent correct to 0 percent correct.

Current Army policy is that the SQT total score scale is divided into three categories. The higher passing score, called the Qualification Score, which is set at 80 percent of the scorable units correct, determines eligibility for award of the next higher skill level, and therefore eligibility for promotion. Only persons with the appropriate skill level are eligible for promotion. The lower passing score, called the Verification Score, which is set at 60 percent of the scorable units correct, determines eligibility to retain the current skill level. Soldiers with SQT scores below 60 percent correct may be reclassified to another MOS.

If SQT scores are also used to rank order soldiers, then in most cases the criterion-referenced power of the tests will be reduced or lost entirely. The following cases illustrate this point:

1. Case 1. If the quotas and number of eligible soldiers are the same, then the decisions of whether to promote, based on the hurdle, and when to promote, based on rank order, have the same boundaries and there is no conflict between quotas and standards.

2. Case 2. If the number of eligibles is less than the quota and the standards are waived until the quotas are met, then the rank ordering would be used to decide both whether and when to promote. Waiving standards could be equivalent to rank ordering. If the standards are waived one unit at a time until the quotas are satisfied, then the effect is to rank order with no regard to prerequisites. The waiving could be done in larger units, say, from 80 correct to 60 correct, and then making the decision of when to promote on the basis of other factors. How the waiving is accomplished and how the tradeoff between standards and quotas is achieved are policy decisions. Waiving standards forces an explicit decision about the tradeoff, whereas the pure rank ordering approach ignores any consideration of standards. On the other hand, if standards are not waived, then the rank ordering would be used only to decide when to promote. In this case the quotas would be waived in favor of increased quality.

3. Case 3. If the number of eligibles is greater than the quota, then, depending on how the pool of eligibles becomes replenished, the prerequisite standards may have varied meaning. If the pool of eligibles is always larger than the quota, then some soldiers near the cutting score may not be reached and consequently not promoted. If the pool is exhausted before new soldiers are added, then these soldiers are assured eventual promotion, and new soldiers who become eligible are placed into a hold category until the original pool is exhausted. If the new eligible soldiers are immediately added to the pool, then there is no assurance that the remaining eligible soldiers from the original pool will be promoted even though they surpassed the prerequisite standards.

The main point about hurdles vs. rank ordering is that the criterion-referenced standards may be lost to the rank order unless explicit decisions are made to retain the standards. Rank ordering lends itself so easily to satisfying quotas that performance standards may be readily bypassed. The ability to obtain objective standards of job performance has profound impact on how personnel decisions can be made. Personnel managers have a choice between using a priori derived standards, independent of the population taking the test, and using quotas derived independent of the content of the test. The traditional solution to personnel decisions is to establish quotas, and then to select individuals until the quotas are satisfied.

According to the criterion-referenced test model, levels of performance within a proficiency category are not discriminated because the criterion levels are the only points of interest. Continuous scores are available, however, and they can be used for rank ordering soldiers. Because SQTs can be scored either in terms of performance categories or as continuous scores, explicit decisions can be made about which methods or combination of methods to use, and how the scores will be used in personnel decisions.

As a minimum, SQTs are used to set prerequisites for promotion. As described above, the prerequisite score is waived to meet quotas if such a policy decision is made. An immediate question is whether SQT scores should be used to rank order the pool of soldiers eligible for promotion. To oversimplify the question: SQTs are now used to determine whether to promote. The question of when to promote can also be answered on the basis of SQT scores, or can be based on other factors. (Other factors besides SQT scores do affect promotability, but the oversimplified version puts the issue in stark relief.)

An unfortunate consequence of using quotas is that performance standards, which may be used in delineating a quota lim t for one particular point in time, may not be entirely relevant when applied in another situation. If, for example, the top 50 percent in a job is eligible for promotion, the job performance of the eligible group will vary as the soldiers change over the years, as the effectiveness of the training programs changes, or as the relationship between test content and job requirements changes over time.

A major breakthrough resulting from criterion-referenced SQTs is the availability of objective information about job competence that can be included in making personnel decisions. Level of job performance measured by these tests provides an absolute indication of proficiency that remains relatively constant as long as jobs remain defined by existing Soldier's Manuals. Performance standards for personnel decisions can be specified in terms of the percentage of job tasks soldiers can perform. These standards are external to the test and, therefore, more powerful statements can be made about the groups that are eligible to be selected in or out.

Quotas for personnel actions, such as promotion or attendance at a school, are likely to remain a driving force for personnel management in the foreseeable future. Rarely, if ever, will the number of soldiers eligible for a personnel action, based on performance standards, be the same as the required quota. Some adjustment to the quotas or performance standards, or both, generally will be required. If quotas are given top priority, then standards are waived; conversely, if performance is given top priority, then quotas are waived. If both quotas and performance are waived, say within some prestablished bounds, then a tradeoff between quality and quantity can be established.

Decision rules about quality vs. quantity can be explicitly stated. If performance standards are waived, there is a cost in terms of lowered individual performance (quality) in order to obtain sufficient numbers (quantity). If quotas are waived, there is a gain in individual performance (quality), but insufficient numbers (quantity) are obtained. By assigning values to units of performance and shortfalls, the tradeoff between quantity and quality can be calculated. Again, the tests do not dictate policy about quantity or quality, but they support decision rules and permit operations not possible without them.

The situation becomes more complex when personnel decisions are not based exclusively on test scores, but rather include test scores as one factor in a composite score. Army personnel actions generally have been based on a composite score, which is characterized as the whole-man concept. The composites may be governed by explicit rules to provide objective indices, or the variables may be combined in a subjective manner by the decision makers. An example of explicit rules governing the combination of factors is Enlisted Evaluation Scores based on a weighting of MOS test scores and Enlisted Evaluation Report scores; another example is the determination of whether a soldier meets the prerequisites for a particular job training course, in which aptitude area scores, physical profile, and perhaps prior training may be considered. An example of subjective combination of factors is the process followed by a typical selection board that interviews soldiers, examines their records, and then arrives at a collective decision.

Criterion-referenced standards require the use of explicit rules for setting the minimum levels of qualification. If the process of combining scores for the qualified group is objective, explicit weights are assigned to each variable, and the contribution of each variable to the component score can be specified.

The assigned weights and the actual weights may or may not be the same. The actual weight of a factor is determined largely by the variability or range of scores for that factor. If the range is small, the effect is to add a virtual constant value to each individual's score, regardless of assigned weight, and the small differences can have only a small effect on the final rank ordering of the soldiers. If the combining is based on subjective judgment, then the weighting of the variables cannot be explicated. In either case, an important consideration is how the minimum qualifications are treated in determining eligibility for a personnel action. If the standards do serve to categorize soldiers into qualified and nonqualified groups and the qualified group is then given the favorable treatment while the nonqualified group is excluded from consideration, then the criterion-referenced standards are operative. If, however, the minimum standards can be waived, then the subjective process may easily ignore the standards, and the net effect may be to lose the power that inheres in criterion-referenced standards.

The process of combining scores may also be based on successive hurdles. The use of successive hurdles for combining scores virtually assures that standards will be maintained. Establishment of the minimum levels of qualifications requires explicit decisions, and any waiving then must also be explicit. An example of multiple hurdles is the determination of eligibility for entrance in a job training course. A minimum aptitude area score is set, and other minimum prerequisites may also be included in the decision, such as physical profiles, prior military job training, and high school courses completed. Not all eligible persons enter a course, but unqualified persons are excluded unless a specific waiver is applied. The use of hurdles is compatible with criterion-referenced standards.

SQTs, because of their criterion-referenced properties, permit basing personnel decisions on objective performance standards. As has been mentioned, technical feasibility does not necessarily dictate policy, and therefore personnel decisions need not be based on performance standards. However, since the possibility exists, rational evaluation of the costs and benefits in changing to new personnel policies can now be accomplished by decision makers.

MAJOR FUTURE PERFORMANCE MEASUREMENT R&D PROBLEMS

The Skill Qualification Testing program is a bold venture into performance measurement. In this paper some of the potential uses and benefits were discussed. The speed and urgency with which the program was developed in the Army catapulted hidden research and methodology issues into the open. As the system is used and spreads within the Army, data will become available for the first time on a scale sufficient to address questions and problems which are indubitably central to a new and radically different approach to performance measurement. Some of the urgent issues which are within the Army's mission and near future plans are:

1. Establish quantitative, functional relationships among performance measurement, job performance, and mission or system effectivenss.

2. Adapt or invent pertinent metrics and statistics for determining reliability and validity of performance measures.

3. Develop techniques for job/task analysis specification.

4. Derive criteria for determining task and job criticality.

5. Build a job performance matrix to relate job tasks and behavioral actions required for performance.

6. Establish scaling methods for determining equivalence of tasks or jobs.

7. Devise valid and economic means for performance sampling.

8. Develop procedures for effectively generating performance measures from critical job specifications.

9. Determine optimal performance measure fidelity requirements.

10. Standardize observational job performance evaluation.

11. Design appropriate inter-rater reliability measurement techniques and standards.

12. Design and evaluate performance test rating and scoring methods.

13. Explore applicability of automated test administration technology.

In a word, invent and reduce to practice a new technology for criterion-referenced performance measurement.

ABOUT THE AUTHOR

Milton S. Katz is presently Technical Area Chief, Individual Training and Skill Evaluation, for the U. S. Army Research Institute. He received a Ph.D. from the University of Rochester in 1959 and subsequently has performed in a variety of technical and management positions. As a physiological-experimental psychologist and human factors scientist he was employed by the U. S. Naval Medical Research Laboratory, the University of Rochester, American Institutes for Research and General Electric. Since September, 1961 he has served as Head, Communications Psychology Division, U. S. Naval Training Device Center, Deputy Associate Director for the Job Corps, Office of Economic Opportunity, vice president and director of program development for Responsive Environments Corporation, and consultant in Interactive Television for the MITRE Corporation.

PLANNING FOR AIRCREW PERFORMANCE MEASUREMENT R&D: U. S. AIR FORCE

Wayne L. Waag
Air Force Human Resources Laboratory
Williams Air Force Base, Arizona

Patricia A. Knoop
Air Force Human Resources Laboratory
Wright-Patterson Air Force Base, Ohio

ABSTRACT

Considerable effort has been and will be expended toward the development of aircrew performance measurement systems. To date, most work has focused on the development of measures for use in ongoing research programs. The Aerospace Medical Research Laboratory has developed measures reflecting human response characteristics which can be used for the evaluation of advanced weapons systems. The Human Resources Laboratory has focused efforts on developing measures of aircrew proficiency which could be used in both ground-based and airborne environments. A major problem which remains unresolved concerns measurement system validation. At present there exists no standardized criteria and procedures for determining the validity of objectively derived measures. Future efforts will focus on the development of measures which reflect the control strategy of pilots. It is anticipated that such models would be useful for future simulation research. A major effort is also planned for the development and implementation of objective flight simulator measurement systems for use in operational training. It is expected that these efforts will lead to the development of operational airborne measurement systems.

INTRODUCTION

The measurement of aircrew performance has become a matter of increased concern to the operational commands within the Air Force as well as the R&D community. Researchers have long realized the central importance of an adequate measurement system and the fact that it represents the foundation of all other R&D efforts. For this reason, considerable effort has been expended toward development and validation of aircrew performance measures in support of other research activities.

It is only recently that the operational environment has come to realize its need for improved performance measurement capabilities. Perhaps the greatest need is to determine if an aircrew can demonstrate accepted levels of proficiency. The assessment of flying proficiency reflects the degree to which the required objectives are met and generally specifies any errors which are committed. The assessment of aircrew proficiency is necessary for initial as well as transition and continuation training. The necessity of insuring that all aircrews meet minimum proficiency standards is of critical importance and cannot be overemphasized. The requirement for proficiency assessment is applicable to performance in ground-based training devices/programs, as well as in the aircraft.

The ability to accurately assess proficiency is also necessary for the proper design of training programs. Since most flying training programs are a mixture of ground-based and airborne instruction, the assessment of proficiency in both domains is necessary for the development of an optimum syllabus. The successful application of the Instructional Systems Development (ISD) approach to training is dependent upon adequate assessment. For this reason, emphasis has been placed on the development of Criterion Referenced Objectives (CROs) in an attempt to standardize the measurement of aircrew performance.

Furthermore, the measurement of aircrew proficiency is necessary to evaluate the effectiveness of training devices and the program within which they are used. The purpose of ground-based training is to enhance performance in the aircraft. The application of the transfer of training methodology as a means of evaluating the effectiveness of ground-based training is dependent upon aircrew proficiency assessment. It follows that the certification of flight simulators in terms of their training effectiveness is also dependent upon an adequate measurement capability. Even in those cases where ground training is carried out on tasks which cannot be tested in the air (for reasons of safety), proficiency measures are required. In this case, performance measures made in the ground-based environment become surrogates for airborne measures.

To date, R&D efforts within the Air Force have focused primarily on the development of performance measurement capabilities for use in ongoing laboratory research programs. Only recently has the emphasis begun to shift toward the development of measurement systems within the operational environment. The majority of measurement research within the Air Force has been accomplished by two laboratories, the Aerospace Medical Research Laboratory (AMRL) and the Air Force Human Resources Laboratory (AFHRL). AMRL has emphasized the development of measures which reflect the effects on advanced weapons systems of human control and response characteristics. On the other hand, AFHRL has concentrated its efforts on the development of performance measurement techniques for the evaluation of aircrew performance. In the following sections, major research thrusts will be discussed and plans for future R&D efforts outlined.

Human Response Measurement for Predicting Weapon System Effectiveness

A continuing area of Air Force emphasis in performance measurement research has been to develop methods of predicting the effectiveness of advanced weapons systems by analyzing their interaction with human response and control characteristics. Work in this area is conducted principally by the Aerospace Medical Research Laboratory (AMRL) and includes the use of ground-based systems such as a centrifuge-based Dynamic Environment Simulator, roll and multi-axis tracking simulators, and other specialized tracking devices for measuring and predicting human response in various weapons system configurations.

Both conventional measures, such as RMS tracking error, and more advanced measures based on describing function and optimal control models of the operator have been developed and applied for this purpose. An example of the former is a study to investigate human capabilities and potential problems in controlling vectored [illegible] fighters in which lateral motion is possible independent of other motions characteristically associated with conventional maneuvering (Loose, McElreath, & Potor, 1976). Measures of vertical and lateral acceleration and tracking [illegible]rs and error rates were computed on an air-to-air gunnery task in which

direct side force control was implemented. Results on such aspects as the probable amount of training required and the effects of various load factors on performance with this type of control were compiled for consideration in aircraft design efforts.

Although conventional measures are useful for supporting some work of this type, they are often not sufficiently sensitive to reveal information about the source and cause of many performance differences. For this reason, AMRL's work in performance measurement has included applications of engineering models of the human in hopes of characterizing subtle aspects of control characteristics that escape detection using conventional measures. For example, both the optimal control and describing function models are being applied in ongoing studies to investigate the information provided to an operator by various motion cues (Junker and Price, 1976; Levison, 1976). Human performance parameters such as leads, lags, and control gains can be derived through these models to supplement the data provided by standard measures. These modeling approaches are also being applied to the evaluation of a high acceleration seat for air combat. In this case, however, the application is pushing the state-of-the-art because the models have not yet been fully adapted to use in the multi-input, multi-output environment of air combat. In accordance with this, future research thrusts of AMRL will include adaptations of engineering models to more types of situations typically encountered in the real world.

Performance Measurement Techniques for Assessing Aircrew Proficiency

The objective of past and current research efforts at the Human Resources Laboratory has been to develop sensitive, accurate, and reliable techniques for the assessment of aircrew proficiency in both ground-based and airborne environments. To date, the primary emphasis has been directed toward the development of a measurement system for use as a tool in the ongoing laboratory research program. These efforts are summarized in the following sections.

Methodological Considerations. During the late 60's and early 70's, attention focused on the development of improved, systemized methods of deriving valid performance measures for aircrew assessment. Without such methods, researchers tasked with developing measurement systems are forced to either use measures previously developed for other applications or else pursue a lengthy process involving several iterations of manual selection, test, re-selection, and retest of candidate measures. The first method cannot be guaranteed to result in optimal or even satisfactory measures. The second method is equally undesirable because its success is overly dependent on the ingenuity and patience of the particular investigator performing the work. A systematic methodology for developing and testing performance measures would both improve and standardize the quality of measures for future applications.

Two approaches to measurement development and validation were investigated. They differ primarily in terms of the order in which various validation tests are performed and the allocation of research tasks to man and computer. The research tasks associated with deriving measures include: (1) selecting the specific measures to be explored, (2) assuring their content validity, and (3) testing them empirically for other types of validity. The first task can be performed by either man or computer. The second is most efficiently performed by man, since it requires standardized processing of large quantities of data.

In the first approach--termed the empirical approach--the computer is assigned the task of generating candidate measures and performing empirical validation tests, followed by manual analysis of results and assurance of the measures' content validity. In the second--termed the analytical approach--man is assigned the job of selecting candidate measures and assuring their content validity, followed by computer tests of the measures for various types of empirical validity. The first approach (Connelly et al., 1969, 1971, 1974 (a)) places the greatest research load on the computer and has the advantages of assuring examination of a broad spectrum of measures and of being potentially applicable across diverse performance tasks. Its disadvantages stem primarily from the quantity of data to be processed and the attendant need for efficient computer algorithms to handle complex tasks never before implemented on a computer. The second approach (Connelly et al., 1974 (b)) places the greatest research load on the man, and has the obvious advantages of traditionality and apparent simplicity. It is subject to limitations on the number and types of measures that can be identified and explored, as constrained by the ingenuity and available time of the researcher and the number of measurement problems to be addressed.

Both approaches were formalized and preliminary evaluations performed. Although airborne data collection problems interfered with the completion of testing, several distinct accomplishments emerged from these preliminary methodology studies. With the first approach, which is based on computer generation of candidate measures, it was shown that resulting measures tend to be considerably more diagnostic than the traditional error and summary measures and that automatic "weeding out" of measures lacking empirical validity is feasible. However, a great deal of researcher interaction with the computerized process was necessary to make it work due to the newness of the method and lack of several required computer algorithms. The second approach resulted in a formal method of analyzing flight maneuvers to identify various segments in which the pilot's primary control functions involve well-defined and easily measured variables. This led directly to identification of the types of measures applicable to assessing performance in each maneuver segment. The basic problem with this approach, that is, its dependency on the resourcefulness of the researcher, was not resolved. However, an efficient method was developed for analyzing maneuvers to facilitate to the greatest extent possible the identification of meaningful candidate measures.

Automated Simulator Measurement. One of the initial problems was to limit the scope of measurement development activities and to concentrate available resources on one or more selected areas. The domain of aircrew proficiency measurement is simply too great to address the entire problem. In selecting a specific area for measurement development, a number of factors were considered. First, the device for which an objective performance measurement system was required was the Advanced Simulator for Pilot Training (ASPT). The ASPT simulates the Air Force's primary jet trainer, the T-37B. Since the device represents a full-mission trainer, it seemed necessary that a measurement capability be developed for representative tasks for all phases of T-37 training.

The ASPT was designed to be used as a training research tool. Within the training environment, the primary need for measurement is to assess proficiency. In keeping with the Instructional Systems Development (ISD) approach to training, tasks to be learned are defined and evaluated according to their behavioral or criterion-referenced objectives. Since flying tasks are goal-directed, proficiency should be assessed in terms of the degree to which the specific behavioral objectives are attained. Consequently, the criterion-referenced approach was

adopted for subsequent measurement development efforts for the ASPT. In other words, the second approach to measurement development described previously was pursued.

Despite the emphasis on the analytical development of criterion-referenced measures, other types of measures seemed to be desirable from a research standpoint. The most appropriate types of measures for different skill/experience level combinations are still unknown. For example, criterion-referenced measures may be sufficient to assess the proficiency level of students initially transitioning into an aircraft, but inappropriate for detecting skill degradation in experienced pilots. Consequently, measures of smoothness as reflected by aircraft rates and accelerations and measures of control input behavior were included in the measurement scenarios.

One of the unique capabilities of the ASPT was the preprogramming system whereby sets of computer instructions could be entered and executed in real-time. This feature made possible the development of measurement scenarios which could be used in real-time. Such a system provides immediate knowledge of results as well as eliminates the need to store data on some medium and later analyze it using an off-line computer program. Despite these advantages, the real-time approach demanded that careful attention be given to the precise definition of the measurement scenarios. Not only must the parameters be specified, but the rules which determine when the measurements are to begin, how the parameter data are to be summarized, and when the measurements are to be terminated must also be defined.

The flight parameters to be sampled were of two types--those reflecting the state of the aircraft and those reflecting the control inputs of the pilot. The adopted approach assumed that superior pilot performance could be defined by: (1) accomplishing specific task objectives as defined by the aircraft state parameters; (2) avoiding excessive rates and accelerations so that the task is executed smoothly; and (3) accomplishing these objectives with a minimum degree of effort. It should be emphasized that these represented hypotheses about pilot proficiency which would be subject to experimental verification.

At the outset, the decision was made to begin with the simpler flight maneuvers wherein the criterion-referenced objectives could be most easily defined. If the approach proved feasible, then the more complex tasks would be attacked. The first scenario to be developed and implemented was for straight and level flight. Upon completion, subjects of differing experience levels flew the scenario while being evaluated by experienced instructor pilots. An analysis of the data revealed that: (1) the agreement between raters was high; (2) the objectively derived measures predicted the IP ratings quite well; and (3) the objective measures discriminated between novice and experienced pilots (Waag et al., 1975).

Encouraged by these data, scenarios were developed and implemented for additional instrument maneuvers--airspeed changes, constant airspeed/rate climbs and descents, and the steep turn. At this time, development was limited to the instrument environment since the visual system had not yet been installed on the ASPT. Upon completion of these scenarios, additional validation data were collected from which similar findings emerged. The results of these initial validation attempts supported the three hypotheses regarding superior task performance.

It soon became apparent that a data management system would be required as a result of the large number of measures computed for each scenario. Consequently, a data storage and retrieval system, the Student Data System (SDS), was designed to be used in conjunction with the automated performance measurement system. The SDS enabled the storage of all information provided by each measurement scenario. In addition to the computed measures, identifying information such as student name, instructor name, mission number, environmental conditions, etc., could be entered. A provision was also made for entering instructor ratings as well as commentary upon completion of the maneuver. Retrieval programs were developed to enable the editing and processing of data directly from the student files thus eliminating manual data handling. Since its original development, the SDS has undergone numerous changes and refinements in order to increase its efficiency.

The development and implementation of measurement scenarios for individual flight maneuvers has progressed substantially. To date, an automated measurement capability exists for selected tasks in all phases of T-37 training. These include basic instrument, basic contact, takeoff/approach landings, advanced instrument, aerobatics, and formation flight skills. The major emphasis is currently on measurement system validation and refinement. Validation data except for that previously mentioned, has only been collected in conjunction with and incidental to other research studies. Although the available data suggest the validity of these objectively derived measures, there remains a need for a full-scale evaluation since the available data were collected under a wide variety of conditions. An added problem has been the development of a single score which reflects an overall proficiency level evaluation for each specific flight task. At present, summary scores are provided for each measurement parameter. The data collected in the ongoing system evaluation should provide information necessary to develop the required overall total score.

It should be emphasized that measurement development efforts to date have focused on providing a valid and reliable measurement capability within the research environment. Considerations of measurement requirements within the operational training environment have been limited to specific studies in which operational flight proficiency assessments were required. Despite the availability of objective performance measurement data on the ASPT, the fact remains that transfer of training studies require an assessment of proficiency in the aircraft.

Automated In-Flight Measurement. Measurement of in-flight pilot performance is usually accomplished by an instructor pilot who applies a subjective rating scale and places the student in one of several skill categories. Subjective rating is largely a matter of judgment and is subject to many sources of unreliability and invalidity. It also places an unnecessary burden on the instructor who must apply it in-flight and provides no way of assessing solo performance of students or of pilots transitioning to single seat aircraft. Another major thrust of Air Force research in the early 70's was to develop improved and standardized methods of in-flight performance measurement that would also free the instructor pilot from rating tasks that detract from his attention to instruction and safety.

The particular problem to which these studies were addressed was T-37 pilot performance measurement in the Undergraduate Pilot Training (UPT) program. The approach was to develop and implement instrumentation for recording T-37 flight

data; and to develop technology and computer software for automatically measuring pilot performance using the recorded data. This feasibility study (Knoop and Welde, 1973) involved an extensive aircraft instrumentation and flight test program. Twenty-four flight variables including aircraft body axis angles and rates, altitude, heading, engine data, and control surface deflections, were recorded at rates of 10 or 100 per second using appropriate sensors and a data acquisition system which encoded the data in digital form on magnetic tape. Computer programs were written to automatically compute a number of measures selected on the basis of their content validity.

Data on two aerobatic maneuvers, the lazy-8 and barrel roll, were collected using both instructor and student pilots. Various validation tests were applied to evaluate the candidate measures, and several measures were selected and combined to form summary measures for the two maneuvers. Debriefing charts were designed for use in pictorially describing the performance of each maneuver and conveying measures and diagnostic comments to instructors and students.

This study was highly successful in demonstrating the advantages of in-flight data recording and developing candidate methods of automated measurement based on the recorded data. It also demonstrated the type of data required for thorough validation-testing of measures and provided a basis for scoping future efforts relying on calibration, smoothing, and processing of in-flight data. It inspired a follow-on study, conducted in 1973-74, in which the aircraft instrumentation was improved to extend the recording range and reliability for several variables and stick-force sensors and a video recorder were added. The instrumented aircraft has been used by HRL in several measurement and flight simulation research programs. The most recent study collected equivalent data in the ASPT and the aircraft in an attempt to determine the relationship of measured performance in these two environments.

The Problem of Validation. For most complex tasks, there is no single necessary and sufficient test that can be applied to candidate measures to assess their validity. Measures which appear to have concurrent validity may or may not satisfy other validation criteria, depending on the reliability and sensitivity of the metric used as a basis of comparison. Therefore, several validation tests rather than just one must often be designed and applied during the course of a measurement development effort.

One approach to the validation problem and satisfaction of the need for several different tests was developed by the Air Force by viewing validation as a screening process (Connelly et al., 1974 (a)). The first step of the process screens out measures of virtually no utility by applying a general test of content validity. Then three empirical validation tests are applied, each of which is increasingly more stringent. One test assesses the measure's potential contribution to discriminating between performances at opposite ends of the skill continuum (e.g., novice vs. expert pilots). A second test then assesses the measure's functional relationships with concurrent measures of performance (e.g., subjective ratings). Measures that tend to reinforce these concurrent measures either differentially or ordinally, as the case may be, are considered more likely to be valid than those which consistently fail to do so. Finally, a third empirical test assesses the measure's functional relationships with variables such as number of trials and time in training. A measure which demonstrates that learning has occurred from novice to experienced levels of performance would possess a higher likelihood of validity than one which consistently fails to do so.

The problem of validation, already complicated by the lack of a single, necessary and sufficient test, is made even more difficult by a lack of standardization of validation criteria for any one type of test. For example, if a test is to be applied to determine whether or not a measure correctly indicates that learning has occurred, there are no guidelines for judging whether or not the measure in question provides such an indication. Similarly, if a test is to be applied on whether a measure discriminates between two performance extremes, there is no single, accepted method of judging whether or not that discrimination exists and is sufficiently reliable for measurement validation.

Lack of standardized validation tests and systematic methods of applying them to quantitative performance measures is a serious problem that deserves much more attention than it has received. However, a more serious problem arises out of the way in which validation tests are typically conducted. The usual procedure is to demonstrate how the measures in question are related to some existing criterion of performance. Since an existing criterion is used as a basis for validation, it is quite difficult to judge whether or not an improvement has been made. This problem adds to the difficulty of developing improved measurement techniques.

Measurement for Future Simulation Research

As many investigators over the years have observed, one of the first tasks that should be performed in a measurement development program is to define the purpose of the measures that are sought. Until that is done, the problem has not been defined, and an approach cannot be formulated. In the Air Force, there are principally four purposes of aircrew performance measures, and ordinarily one of these purposes is dominant for any given application and drives the attendant measurement research requirements. These are: (1) assessment of current aircrew proficiency to evaluate training progress; (2) prediction of performance on another task or in another machine to identify training requirements or predict mission success; (3) provide learning feedback; and (4) measurement of performance to characterize and identify changes in behavior in support of simulation and training research.

Research devoted to satisfying the first three purposes of measurement systems has far exceeded that devoted to the fourth. In addition, measures that satisfy the first three purposes cannot be assumed to satisfy the fourth. For research applications, measures are required which are highly sensitive to changes in skill, which detect and can be used to identify changes in behavior, and which characterize the manner in which and the extent to which various cues are perceived and used. Only this type of measure can reveal the true effects of different simulator techniques on the behaviors and strategies carried forward for use in the operational aircraft. Research on measures suitable for research applications is sorely in need. It is a well known but often neglected fact that in human learning and performance research, the nature of the proficiency measures used is as important in determining the results obtained as the independent variables employed (Bahrick, Fitts, and Briggs, 1957).

One plan of attack being pursued by HRL is to apply human operator modeling concepts to the problem. Often, the most concise way to represent a set of data is to model the process that generated it. If modeling techniques were applied to human performance measurement, it is conceivable that an optimally concise set of measures could be produced from the model itself. If the model

were carefully formulated and validated, measures derived from it would characterize human behavior rather than the effect of that behavior on system response; and they could be made to include the impact of various cues and the way they are perceived, interpreted, and applied.

A recently completed survey effort conducted by HRL (to be published as an AFHRL Technical Report) revealed that previous human operator modeling efforts have often emphasized matching human response with model response and have failed to insure that the modeling method is in harmony with known human performance characteristics. In other words, the content validity of the model itself has not been assured. In addition, existing models have been oriented toward reproducing the average performance of highly skilled, highly motivated operators without regard to novice or intermediate performers. Finally, models of the past and present have not yet successfully incorporated the use of multiple cues from various sources (visual, proprioceptive, kinesthetic) and have placed too little emphasis on behavior, strategy, and learning and too much on performance as represented, for example, by RMS error. Efforts are underway in HRL to derive a suitable model of control strategy development and learning for use in measurement for simulation research. Although concepts represented by existing models are being applied wherever possible, the bottom line of the work is to assure that known or accepted theories of human performance constitute the major driving force for model development.

Measurement for Future Operational Flight Training Systems

Within the Air Force, there has occurred an increased emphasis on flight simulation training primarily as a result of spiraling aircraft operating costs. In order to document the effectiveness and efficiency of the simulator training syllabus, it is desirable to quantitatively assess aircrew proficiency both in the simulator and the aircraft. Once a measurement capability exists, it is possible to precisely determine the effects of changes in the syllabus of instruction, and, therefore, optimize the flight simulation and aircraft training programs. Such a system would also allow the precise definition of proficiency requirements. In this manner, the flying training manager could more readily control the quality of graduates from his training programs.

The development of objective performance assessment capabilities for operational flight training systems will most likely proceed in two phases. The first phase will focus on the implementation of measurement systems within the flight simulation environment while the second will focus on the aircraft. There are definite advantages in accomplishing the initial development efforts in the simulator. Since the simulation environment can be precisely controlled, the development and validation of assessment algorithms can readily be accomplished. Likewise, the flight simulator provides access to all control input and aircraft state parameters whereas the number of parameters sampled in the aircraft environment is usually limited. Selecting the most critical parameters based on simulation data should save time and reduce costs in the development of an airborne measurement system.

Within the near term future, HRL plans to initiate a 5-year program for the development and installation of objective performance assessment systems in selected flight simulators. The program will have three major thrusts. First,

an automated performance measurement system will be developed and implemented for selected simulation systems. Since continuous performance monitoring is impractical, the goal is the development of one or more automated simulator checkrides which could be used in an analogous manner as an aircraft checkride.

A second major thrust will be an evaluation of the utility of automated performance measurement in operational training. First, the correspondence between measured performance in the simulator and aircraft will be determined. The prediction of performance in the aircraft from performance in the simulator is viewed to be a major potential benefit of the program. Second, the measurement system would be evaluated according to the usefulness of the feedback provided to the student and instructor. And third, it would be evaluated according to the usefulness of the information provided to the syllabus designers.

A third major thrust will be the development of a set of functional specifications for the inclusion of a performance measurement capability for future generation devices. It is expected that requirements would also be generated for the development of retrofit capabilities for existing simulation systems.

Several criteria have been established for the selection of simulation systems for which a measurement capability would be developed. First, it is desirable that the systems represent a broad range of aircraft types and missions. Given the limited resources, it is important that the results be as generalizable as possible. Furthermore, the simulation systems, per se, should represent different levels of "state-of-the-art." At least one system should require a retrofit, while another should be able to support the implementation of the measurement system using existing hardware. Where possible, it seemed that the selected systems should build upon existing technology.

The focus of the 5-year program just described is the development of objective proficiency assessment systems for flight simulation. Nevertheless, pilots are trained to fly aircraft, and it is this environment for which objective proficiency assessment techniques are ultimately needed. It is anticipated that the long range performance measurement efforts will be directed toward providing this capability in the aircraft. The successful development and implementation of objective assessment in the simulation environment should pave the way for airborne performance measurement.

REFERENCES

Bahrick, H. P., Fitts, P. M., and Briggs, G. E. Learning Curves--Facts or Artifacts? Psychological Bulletin, 1957, 54, 256-268.

Connelly, E. M., Schuler, A. R., and Knoop, P. A. Study of adaptive mathematical models for deriving automated pilot performance measurement techniques. AFHRL-TR-69-7, Volumes I and II, AD-704597 and AD-704115, AFHRL, Wright-Patterson Air Force Base, Ohio, 1969.

Connelly, E. M., Schuler, A. R., Bourne, F. J., and Knoop, P. A. Application of adaptive mathematical models to a T-37 pilot performance measurement problem. AFHRL-TR-70-45, AD-726632, AFHRL, Wright-Patterson Air Force Base, Ohio, 1971.

Connelly, E. M., Bourne, F. J., Loental, D. G., and Knoop, P. A. Computer-aided techniques for providing operator performance measures. AFHRL, AD-A014-330, Wright-Patterson Air Force Base, Ohio, 1974 (a).

Connelly, E. M., Bourne, F. J., Loental, D. G., Migliaccio, J. S., Burchick, D. A., and Knoop, P. A. Candidate T-37 pilot performance measures for five contact maneuvers. AFHRL-TR-74-78, AFHRL, Wright-Patterson Air Force Base, Ohio, AD-A014331, 1974 (b).

Junker, A. M. and Price, D. Comparison between a peripheral display and motion information on human teaching about the roll axis. The 12th Annual Conference on Manual Control, 1976.

Knoop, P. A. and Welde, W. L. Automated pilot performance assessment in the T-37: A feasibility study. AFHRL-TR-72-6, AFHRL, AD-766446, Wright-Patterson Air Force Base, Ohio, 1973.

Levison, W. H. Use of motion cues in steady-state tracking. The 12th Annual Conference on Manual Control, 1976.

Loose, D. R., McElreath, K. W., and Potor, G. Effects of direct side force control on pilot tracking performance. AMRL-TR-76-87, Wright-Patterson Air Force Base, 1976.

Waag, W. L., Eddowes, E. E., Fuller, J. H., and Fuller, R. R. ASUPT automated objective performance measurement system. AFHRL-TR-75-3, AFHRL, AD-A014799, Williams Air Force Base, Arizona, 1975.

ABOUT THE AUTHORS

Wayne L. Waag, a native-born Texan, received his undergraduate degree from the University of Houston in 1966. He entered the experimental psychology program at Texas Tech University and received the Ph.D. degree in 1971. Upon graduating, he was awarded a National Academy of Science postdoctoral research associateship with tenure at the Naval Aerospace Medical Research Laboratory. In 1973, he accepted a position at the Flying Training Division of the Air Force Human Resources Laboratory. At present, Dr. Waag is the Assistant Chief of the Flying Training Research Branch.

Patricia A. Knoop is a Project Scientist in the advanced Systems Division of the Air Force Human Resources Laboratory. She received an AB Degree in Mathematics in 1962 from MacMurray College, Jacksonville, IL. Since 1963, she has performed research in various areas of flight simulation for crew training, including numerical integration methods, computers and software, instructor/operator station design, advanced instructional features, and performance measurement. In 1975, she received an MS Degree in Computer and Information Science from the Ohio State University. She is a member of IEEE, ACM, and the Human Factors Society, has a commercial Pilot's license with instrument rating, and is an avid bicycling enthusiast.

THE DEVELOPMENT OF A NAVY PERFORMANCE EVALUATION TEST FOR ENVIRONMENTAL RESEARCH (PETER)

Robert S. Kennedy, CDR MSC USN
Head Human Performance Division and Officer-in-Charge
Naval Aerospace Medical Research Laboratory Detachment
and
Alvah C. Bittner, Jr.
Human Factors Engineering Branch
Point Mugu Test Center

ABSTRACT

The basic problem with performance testing in exotic environments is the general unwillingness of investigators to take the time to standardize a test battery. Many other problems exist and are obvious to all who have tried to measure performance under usual and unusual environmental conditions. It is the purpose of this paper to set forth some of the problems that have grown out of our experiences and which we feel have not been extensively commented upon in the research literature, and also to describe our plan for solution.

Preface

The present plan is a simple one: The literature will be searched for human performance tasks which have been shown to degrade under motion (vibration and ship motion), during thermal exposure, and under pressure. The performances that meet these first criteria will be categorized as cognitive (decision making, information processing, judgment), motor (tracking, reaching), etc., and a taxonomy of performances will be developed. Additionally, each performance task will be evaluated in the following way: 20 subjects will be tested 10 times (5 days/week for 2 weeks) to determine three types of reliability: internal consistency, the accuracy and sensitivity to separate individuals, and the stability of this accuracy and sensitivity over repeated testing. Performances on these tasks will be compared to scores on other tests of mental functions. Progress to date will be reported.

The National Aeronautics and Space Administration, the Advanced Research Project Agency, the Navy (via the Office of Naval Research), and the Bureau of Medicine and Surgery have funded several studies (see Kennedy, 1977 for a review) which have nearly all made very similar points regarding the standardization of a performance test battery for assessment of environmental stressors. In the main, test batteries have been proposed, particularly factor analyzed batteries, but rarely have normative data been collected and never have practice effects been studied effectively.

The original title for the present paper was very broad and included all Navy R & D concerning performance. We intend, however, merely to present how the Naval Aerospace Medical Research Laboratory Detachment plans to research the general area, with specific application to our interests in the effects of ship motion or performance. It should be noted that, in addition to the human performance R & D already presented at this symposium by various members of the Navy

Personnel Research and Development Center, complementary programs also exist within the Engineering Psychology Programs of the Office of Naval Research and within the Human Effectiveness Programs of the Naval Medical Research and Development Command.

INTRODUCTION

Casual observation over several years of performance testing and a comprehensive reading of over 400 "human performance studies" in hyperbaria (see Bachrach & Kennedy, 1977, for a review) suggest that there is a need for future studies into the standardization of a human performance test battery.

In our opinion, the persons who initiated the experiments requiring performance testing in exotic environments were generally persons who became involved originally because of a primary interest in the environment rather than in the performance. (Within "environment" we include unusual sensory stimulations, drugs, fatigue, and even learning, as well as motion sickness, hyperbaria, etc.) Thus, we feel that, frequently, several criteria were employed (often trading back and forth among them) in the selection of tasks for inclusion in a battery to be assembled. These criteria have included the following:

1. Literature findings that were recollected, probably because the results of tests were unusual.

2. What colleagues and friends had done.

3. What demonstration experiments were performed in experimental psychology laboratory during their student days.

4. Chapter headings in Woodworth and Schlosberg (1954) and other standard texts.

5. Equipment left behind in the storage room of the laboratory by their predecessors.

6. That which could be quickly and easily assembled from clever ideas, (the so-called toy gadget approach).

7. Stock items from apparatus companies.

8. Logistic limitations forced by the environment or project (e.g., small, inexpensive, no tubes, portable, nonmagnetic, self-scored, no sparks, self-administered, battery powered, and rugged).

9. Similar to the work done by real-world persons.

10. A relatively basic kind of skill is involved; that is, learning theoretically SHOULD be able to be accomplished quickly.

11. Less often, performances could be expected to be disrupted on the task in <u>this</u> environment.

We believe that the criteria listed above have been employed often enough to assemble batteries so that these criteria are worth citing. It should also be noted, however, that, typically, a test battery was generally an ad hoc response to the imminent availability of an environmental condition, whether the environment was a hurricane (Kennedy, Moroney, Bale, Gregoire, & Smith, 1970), a rotating room (Guedry, Kennedy, Harris, & Graybiel, 1964; Fregly & Kennedy, 1965; Kennedy, Tolhurst & Graybiel, 1965), or a deep dive. Thus, long-range planning frequently is not possible. In summary, it is felt that performance test batteries are often assembled for largely practical reasons, on short notice, by persons whose major interest is not performance testing. To alleviate these problems we have combined, in tabular form, what we consider the traditional, important criteria for test construction along with the practical aspects concerning operational performance assessment. These criteria are summarized in Tables 1 - 4. In addition, other problems with performance test battery construction exist.

1. What performance tests are designed to measure

Although this distinction is not generally made, it is implicit that performance testing is undertaken for two main purposes: first, to be able to make some statement about the integrity of the organism, and second, to determine whether an environment interacts with an organism's ability to do a particular kind of work (cf. Table 3). In this paper, the first purpose will be called "CNS status," and the second, "effectiveness of a system's output." Examples of tests designed for the former purpose include reaction time, digit span, tremor, electroencephalogram, speed of tapping, and CFF. Examples of the latter include an underwater pipe puzzle, a sonar monitoring task, Morse code tests, and speech intelligibility tasks. Frequently, both types of tasks are included in a single experiment into the environment's effect on man and without regard to the distinction made above. The advantage of the latter approach is that the system's concept is used and the translation to real-activities is direct. (Also, subject cooperation is usually better.) The disadvantage is that no general principles are adduced and the application of the findings holds only for the stimulus condition employed. For instance, tracking studies with CRT displays have been conducted for many years and very few general rules have resulted (Adams, 1961). The major disadvantage of the first approach (index of an organism's integrity) is that they depend heavily upon the knowledge of the validity of the task. If only face validity is available, other considerations (money, size, apparatus, and availability) must be used to justify inclusion. If face validity is not evident, then justification is very tenuous.

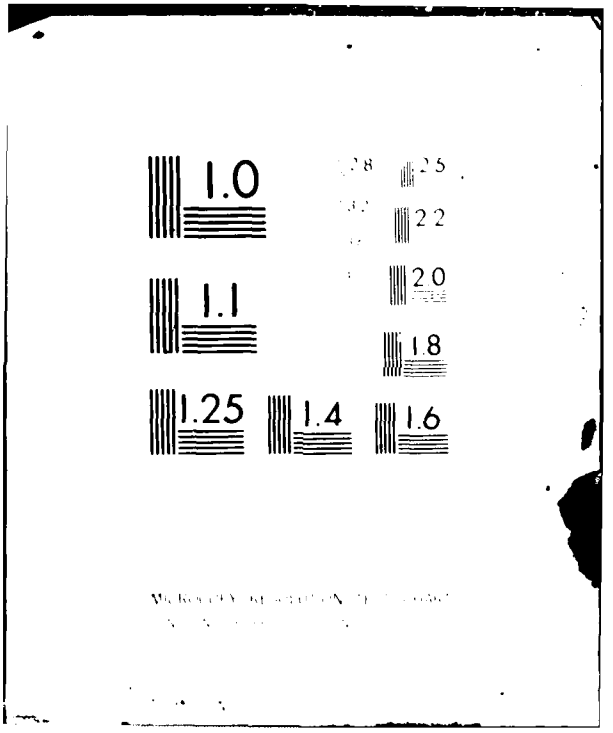
The distinction made between these two strategies is subtle, but it is also real, and its existence complicates the results of many studies. This is chiefly due to the fact that the two approaches require different research philosophies, although the ultimate aim of both approaches is similar: namely, prediction (i.e., an ability to account for 100 percent of the variance).

The first approach comes directly from experimental psychology and usually follows an analysis of variance model. Thus, the numerous tests in a test battery are designed to sample all of the skills (factors) of the organism. The implication is that, if the full range of human abilities is tested, one can generalize the findings and apply them to other circumstances (e.g., subjects, treatments, etc.). This approach depends heavily upon following the principles of test

AD-A116 344 NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 5/9
SYMPOSIUM PROCEEDINGS: PRODUCTIVITY ENHANCEMENT: PERSONNEL PERF--ETC(U)
1977 L T POPE, D MEISTER

UNCLASSIFIED NL

5 OF 5
AD A
116344


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2

construction: (1) norms, (2) reliabilities, (3) validities, (4) factors tested, (5) effects of practice, and (6) individual differences. If all these principles were satisfactorily fulfilled, it would be possible to employ the test in an exotic environment and account for all the main effects of such an environment on human performance. For example, if it were known that hand dynamometry correlated perfectly with all other kinds of voluntary skeletal muscle output, and the Harvard Step Test (Kennedy & Hutchins, 1971) with all cardiac muscle output, then it would not be necessary to use other tests of these functions. The difficulty, of course, is that neither of these tests correlates sufficiently. Additionally, other "more psychomotor" tasks are even less clear-cut with regard to what they are measuring (i.e., validities). However, the problem does not end here. Reliabilities of a test battery--any test battery--are not completely known. No norms (expected values) are available on a sizable population, particularly when practice effects are concerned. However, factor analyses studies (e.g., those of Fleischman) have been completed for some samples.[1]

The second approach is in vogue more now than previously, probably because it emphasizes a systems approach. The statistical model employed is correlation, and in general, single factor studies are conducted. The overall plan is to replicate real-world work and to do it under controlled conditions. The second approach does not depend upon the validity of the task as heavily as the first method, since it, itself, is the work. However, the characteristics of the subjects are critical. It is important, and usually essential, that the subjects be the same kind of people as the real-world workers toward whom the data will be applied. The shortcoming of this strategy is also its chief advantage: the application of the findings from such studies is specific and immediate, but sometimes it is so specific that generalization within the same environment, but with slight differences, may not be possible.

2. Two experimental paradigms

There are two main ways in which to study the effects of the environment on a subject's ability to do work. The first (most often used) uses the subject as his own control and generally follows a pre-, per- and post- paradigm. In the pretest, the subject is practiced on all the tests to be employed in order to arrive at a learning plateau. Then he is placed in the experimental situation to see whether or not it disrupts performance. Posttesting is used to monitor recovery effects, if there are any. There are many problems with this approach. Chiefly, psychomotor performance almost never arrives at a plateau. This is discussed in more detail later in this paper. Asymptotes occasionally are obtained, but these, too, are infrequent. Even on tests where one would expect practice to be accomplished quickly (e.g., reaction time, CFF, tracking visual acuity),[2] the environment itself occasionally causes certain tests to be performed less well while standing during rotation, and is probably also measuring

---

[1] Sinbad (1969) is based on these studies and, when standardized, may be used to obviate some of the problems mentioned above.

[2] The use of signal detection theory (Swet, Tanner, & Birdsall, 1961) as a methodology may be helpful here, but as we all know from the way the 100-yard dash record is continually broken, it is not just a criterion problem. Stated differently, a knowledge of sensory sensitivity, d' (d-prime) separated from the subject's criterion (beta) would refine present knowledge, but d', even carefully and prudently measured, may change with practice.

body sway (Graybiel, Kennedy, Knoblock, Guedry, Mertz, McLeod, Colehour, Miller, & Fregly, 1965). This point will also be discussed later. Post-effects also present difficulties since motivation changes (e.g., end spurt in vigilance) usually attend the imminent completion of an experiment.

The alternative approach: to test "just before" and "just after" the environmental exposure (say a 12-hour overwater ASW flight) has its own problems; namely, the experimenter feels that it is necessary to be aware of the status of the subject during the exposure. If the testing is short (e.g., hand dynamometry), it can be influenced by the bias of a subject and summoning efforts for a "one-shot-deal" so that, often, changes are not obtained even though the subject is frankly tired. If the testing period is long (e.g., treadmill), it can contribute to the fatigue. In addition, lengthy posttests are often unfair to the subject.

3. Assessment of input-integrator-output circuits

The general form of psychological experimentation follows an S-R paradigm, or SOR, where O is for organism (Graham, 1951). Performance testing employs this paradigm particularly when "CNS status" type experiments are conducted. Typically, in these studies the experimenter is mainly interested in whether his treatment (drugs, hypoxia, confinement, magnetic fields) produces any CNS change. So, a stimulus is presented and the output of the organism is monitored for changes. Frequently, however, due account is not taken as to whether the stimulus was adequately received by the receptor (retina, ear, hair cells, etc.) then properly delivered along that nerve pathway; also, whether the output (muscle) pathway is similarly unaffected. For example, during acceleration stress, the lack of oxygen to the retina indicates that signals are not adequately received at the receptor site. This also occurs with the differences obtained in visual performance underwater. The physical conduction of light in air versus water may account for these differences -- most likely the visual signal is just not delivered to the receptor in water as well as in air, so one would not posit CNS changes underwater to account for the poorer visual acuity obtained. At the other end of the nerve-muscle circuit, changes in four-choice reaction time done underwater clearly have the friction of water on the one hand to slow down performance as well as the possible other effects of compression and mixed gases and so, probably, CNS changes cannot adequately be assessed with this task. So, too, past pointing underwater may be different: not because of central involvement, but because of inertial differences on the arm. This is not to imply that such studies should not be undertaken, rather, it behooves the experimenter to indicate where possible which part of the OSR circuit he is testing. Therefore, one must know about the transmission characteristics of light, the dependency of the retina on oxygen, and the viscosity and buoyancy characteristics of water. However, if such tasks are included in batteries that have other tests, (the intention of which is to tap the state of the CNS) when all results are reported together, there is confusion.

It would be useful to other investigators if results of experiments were reported relative to that part of the circuit which is being tested. This cannot be done in all cases, but it is possible to improve present reporting practices. Perhaps if we intellectually remove the known physical environmental effects from the periphery (nerve and muscle), we may be left with the finding that motivation

and the partial pressure of oxygen in the brain are the chief contributors to performance decrement under all conditions. The above criticism does not apply to the "systems output" type of studies which take no position regarding where in the circuit the problem occurs. Rather, their sole purpose is to determine whether an interaction of environmental condition occurs on people doing work. It is proposed that "CNS status" be used as a term to be contracted with "input/output quality" types of studies, whereby the former would deal with throughput changes due to the environment and the latter would address the physical aspects of the environment on man.

4. Practice effects

In a significant but not widely referenced paper, Bradley (1962) reported the persistence of sequence effects during psychomotor testing. Virtually all who study performance over many sessions have obtained similar findings. As was mentioned earlier, the investigator usually performs baseline pretesting before placing the subjects in the environment. Often, many trials are given (in one study, 7 days of testing) in an effort to have performance asymptotic "so that the pimple on the line can be more easily seen."[3] What is usually obtained is the well-known learning curve, which may, but does not always, asymptote. The problem with this approach is obvious, but there is another less obvious problem; that is, performance on a task after many trials is probably no longer an index of the same activity or place in the CNS that it was initially.

Studies by Ades and Raab, 1949, on the Kluver Bucy Syndrome (cited in Bachrach and Kennedy, 1977) illustrate the latter point where animals with certain portions of their brains removed were able to perform a visual discrimination task about as well as unoperated animals; however a simularly operated group was never able to learn this task.

Moreover, it is well known from the learning literature that, with extended practice, subjects overlearn, and when something is overlearned, it becomes more resistant to extinction. Therefore, for performance testing in exotic environments, if intensive practice is given on the tests prior to their use in the experimental environment, two factors appear inevitable: (1) the work is not an index of what it was at first, and (2) disruption of performance becomes very difficult. An example of this is as follows: move the index (first) and ring (third) fingers preferred hand together with the palms resting on a flat surface. Then move the second and fourth fingers together. Then, alternate 1 and 3, then 2 and 4, etc. Everyone can do this work, but it requires far more concentration for the average person than for a person who frequently plays the piano. The investigators believe that control for this activity is exerted high in the cortex for nonpianists, but has perhaps been shunted to a lower center in the CNS in practiced pianists. If the above is similar to what occurs in performance testing studies, the implications are obvious.

Because of the problems listed above, the following approach is planned: We feel that the approach is innovative, but it will draw heavily on the research literature for the initial selection of tests to be included for further study.

---

[3]Radloff, 1971, personal communication.

Those tests will be selected from the literature that meet criteria in one of the following areas: (1) demonstrated sensitivity to either thermal, motion, or hyperbaric environments by exhibiting degraded performances, (2) diagnostic capability (i.e., brain-damaged individuals have been found to perform differently from a normal population), and (3) measurement capability of a parameter of human information processing. After initial selection of the tests, the most promising will be subjected to further tests. The test and equipment attributes of each test will be viewed from the standpoint of the following factors ranked in general order of importance: (1) reliability (e.g., test-retest, alternate form, between and within administrations), (2) validity (e.g., predictive, content, construct, diagnostic-concurrent, fact), (3) other practical test factors (range of capability levels covered, sensitivity, transportability, efficiency), (4) equipment factors (e.g., availability, equipment reliability, transformability, safety, economy). Those tests that demonstrate a high level of adequacy on the above criteria will comprise an experimental battery. Performances on this battery will be compared to performances on a factor pure (e.g., Sinbad) battery to determine uniqueness of factors. Paper and pencil tests of cognitive functions (e.g., Bender-Gestalt, Guillford-Zimmerman) as well as well-standardized intelligence tests (e.g., Wais, Ravens, Stanford-Binet, Reitan, Halstead, Wunderlich) will be administered to this same population to further delineate and validate the factors obtained.

The first test that we have selected for further study is the so-called Beeper reviewed by Kennedy and Bruns (1975). The reasons for selecting this test originate partly from the literature review and partly from the study of acceleration stress by the NAS/NRC Committee on Bio-Astronautics, who convened a working group headed by Robert Galambos to discuss and report on principles and problems of performance testing. Using criteria based largely on earlier suggestions of Broadbent (1953), a performance test battery was proposed that would have general and specific applications.

We looked into Broadbent's report for ideas relative to the common problems of motion and acceleration stress and of exotic environments in general. Recommendations were also included for the use of tasks which are: "(a) work paced; (b) require vigilance; (c) over a long period of time; and (d) during which there is uncertainty in the stimulus display" (p. 22):

1. Laboratory norms on six different versions of this task for each of the approximately 100 college graduate males are available, as well as relationships to personality and other subject variables (e.g., hours of sleep) for these persons.

2. Neurophysiological correlates (vestibular nystagmus) of performance were shown.

3. Practice effects appear small on the three-channel auditory version and are known for the three-channel visual version.

4. The test can be group-administered.

5. It is relatively simple and inexpensive to construct.

6. There are many possibilities for constructing alternate forms.

7. Task difficulty can be controlled largely by instructions.

8. Latency of response within broad limits (namely, 1-2 seconds) is generally not a factor and so the task can appropriately be used even when environmental variables can interact physically with response speed (e.g., underwater).

9. Stimulus recording is binary and therefore is mechanically simple. Further, the regularity of the stimuli makes a scoring relatively easy and relatively independent of where on the magnetic tape a session begins.

10. Proportion measures are essentially linear (R .95) with absolute measures (namely, hits) and, therefore, direct comparisons can be made over different tasks.

11. Unlike many other vigilance tasks, many signals and responses occur and so individual time-line analyses are possible.

12. The results suggest that performance on forms of this task may be age-related.

The approach we have utilized includes the daily administration (15 minutes) of the Beeper for 2 weeks to study the reliability of the test in three ways: internal consistency, the accuracy and sensitivity to separate individuals, and stability of this accuracy and sensitivity over repeated testings.

We feel that this approach will serve as a model for future tasks to be included in our battery. At this writing, data are being collected, however the study is not completed. These results should be available at the meeting in October.

Table 1

Equipment Factors

Factors	Definition	References	Comments
Availability	Equipment software and hardware for presenting tasks, receiving responses, recording, scoring and integrating should be acquirable without excessive delays.	Alluisi (1967, 1969); Reilley & Cameron (1968); Kennedy (1971); Theologus et al. (1973)	Rose (1974) has suggested paradigm "reproducibility" (frequency with which a task has been studied) as a criteria for selection. Certainly selection of tasks most readily available in psychological laboratories would insure maximum cross laboratory availability. Some paper-and-pencil tasks rate high on this factor.
Equipment Reliability	Equipment software and hardware must be sufficiently reliable to permit sustained use for lengthy durations, i.e., have a high expected "mean time between failure." (MTBF)	Alluisi (1967, 1969); Theologus et al. (1973)	A method of checking hard and software states--proper or improper functioning--is a necessity for PTB tasks.
Transformability	Tasks can be adapted for administration in various environments of interest without seriously altering measurement capability.	Reilly & Cameron (1968)	Environments of interest could include "shirt sleeve laboratory," exotic environs (e.g., underwater), or field conditions. Portability (Rose, 1974) and potential for group administration (Kennedy, 1971) are valued elements.
Safety	Equipment should not present a potential health or safety hazard to subjects, and equipment must not be vulnerable to damage by stressed subjects.	Theologus et al. (1973)	This is the most important feature of any battery.
Economy	Costs for acquisition of equipment hard and software, administration, scoring, interpretation and maintenance should be reasonable.	Alluisi (1967, 1969); Reilley & Cameron (1968); Kennedy (1971); Theologus et al. (1973)	Temporal and monetary costs are important, albeit able to be traded off. Equipment based batteries have not been extensively applied or developed because of costs being excessive. Less expensive and sophisticated batteries would encourage standardizations of tasks in the literature.

Table 2

Reliability Factors

Factor	Definition	References	Comments
Test-Retest Reliability	Correlation established by administration of the same test to the same individuals on two different occasions.	Alluisi (1969) Grodsky (1967) Kennedy (1971) Theologus et al. (1973)	Experimenters using PTBs frequently administer the same task to subjects a large number of times. This has been shown in the literature to frequently result in changes in the nature of what is being measured and low correlations between early and later trials (cf., Woodrow, 1938 a & b; Fleishman and Hempel, 1955; Parker & Fleishman, 1960; Parker & Fleishman, 1961; Parker, 1964). Test-Retest reliabilities need to be determined over numbers of trials task will be administered.
Alternate-Form Reliability	Correlation established by administration of two "equivalent forms" of the same test (measuring same aspect of performance but of different questions or items) on two different occasions.	Theologus et al. (1973) Teichner (1974)	Alternate form reliability is appropriate for tasks with elements which lose potency when exposed to subjects. Also note that comment for Test-Retest applies with alternate forms which are employed over substantial numbers of trials.
Internal Consistency Reliability	Correlation estimate of the homogenity of a task's item scores established on a group of individuals on one occasion.	Theologus et al. (1973)	Note comment for Test-Retest has implication for internal consistency estimates a point delineated by Thorndike (1949).
Between Test Administrators Reliability	Correlation established by administration and scoring of the same or equivalent form of a task to the same individuals by two administrators.	Teichner (1974)	This reliability has not been of interest in most developments of PTBs, although the "experimenter effect" has a long history in experimental research (cf., Rosenthal, 1961).
Within Test Administrators Reliability	Correlation established by administration and scoring of the same or equivalent form of a task by the same administrator on two different occasions.	Teichner (1974)	This is the special case under which most Test-Retest and alternate forms reliabilities are established.

Table 3

Validity Factors

Factor	Definition	References	Comments
Predictive Validity	Correlation between operator performance on a task (or tasks) and future criterion performance or status.	Alluisi (1967, 1969) Grodsky (1967) Theologus et al. (1973)	"Real world" performance is a concern of experimenters who optimize this criteria vs. diagnosis of performance which is concerned with concurrent diagnostic validity.
Concurrent (Diagnostic) Validity	Correlation between test score and a diagnostic criterion status obtained at approximately the same time.	Teichner (1974)	Teichner (1976) stresses diagnostic aspects or tasks for assessment of subjects internal status (e.g., Is a nervous system dysfunction present?)
Content Validity	Extent to which a task or task battery covers a representative sample of the behavior domains to be measured.	Alluisi (1967, 1969) Reilley & Cameron (1968)	Related to content validity are the concepts of battery "generality" or "comprehensiveness" given as criteria by Theologus et al. (1973), Rose (1974) and Teichner (1974). These concepts stress that a battery should encompass as many critical aspects as possible while minimizing redundancy.
Construct Validity	Extent to which a test may be said to measure a "theoretical construct" or trait where theoretical construct or trait is established by convergence of information from a variety of sources.	Theologus et al. (1973) Rose (1974)	Rose (1974) has particularly stressed this concept by his emphasis on well used "paradigms" from experimental psychology with correlational and factor analysis as methods of convergence.
Factor Validity	Extent to which factor analysis indicates task as identifying or correlating with a factor.	Reilly & Cameron (1968) Theologus et al. (1973) Rose (1974)	Factor analysis has additional use in the assessment of the amount of redundancy in a battery.

Table 3 (Cont.)

Validity Factors

Factor	Definition	References	Comments
Face Validity	Extent to which test "looks valid" to subjects who take it, experimenters, or other observers	Alluisi (1967, 1969) Grodsky (1967) Reilly & Cameron (1968) Theologus et al. (1973)	Alluisi (1967, 1969) and Grodsky (1967) both stress need of face validity to insure subjects feel tasks are relevant and are motivated. Theologus et al. (1973), however, stresses the need of "...face validity to permit subjective generalization of effects...to the effects... on a 'real world' task..." Attempts to measure task face validity have not been reported. Briefing on importance of tasks vs. "face validity" method of motivating subjects have not been reported in literature although used as research strategy (e.g., by Cross & Bittner, 1969).

Table 4

Ability Range, Sensitivity, Trainability and Efficiency Factors

Factor	Definition	References	Comments
Range of Ability Levels Covered	Extent to which differing subject populations (varying in background, developmental level, training, etc.) can be tested.	Alluisi (1967, 1969) Reilly & Cameron (1968) Teichner (1974)	Although pointed out as important, this factor has not been given much study.
Sensitivity	Extent to which test reflects effects of conditions of study.	Alluisi (1967, 1969) Grodsky (1967) Reilly & Cameron (1968) Theologus et al. (1973) Rose (1974) Teichner (1974)	Alluisi (1967, 1969), Grodsky (1967), and Theologus et al. (1973) emphasize sensitivity to effects only to magnitude experienced in operational situation. Reilly & Cameron (1968) define sensitivity as extent to which conditions are likely to influence performance. Teichner (1974), however, discusses it in terms of quickness of detecting dysfunctions. Sensitivity $S = (M_1 - M_2)/SD_1^2 + SD_2^2)$, where M_1 and SD_1^2 are the mean and variance under condition "i" appears a more useful metric for purposes such as Teichner (1974).
Trainability	Asymptotic levels of performance should be attainable with the selected tasks after a minimum of training except where tasks are selected to measure changes in this function per se.	Alluisi (1967, 1969) Kennedy (1971) Theologus et al. (1973) Rose (1974)	To date studies to insure "asymptotic levels of performance" have not been accomplished for non-leanring tasks. Development of tasks to measure different types of learning per se is very lacking though, as Teichner (1974) point out, the most sensitive test will have "minimal practice" as a characteristic. Learning tasks appear to have high potential for future PTBs and should be more fully studied.
Efficiency	Importance of test's contribution with respect to cost, time and effort of implementation.	Reilly & Cameron (1968) Theologus et al. (1973) Teichner (1974)	Contributions of tasks in terms of cost, time and effort appear to be accomplishable by appropriate analysis. The reliability of a task for one minute of study (r_{11}), for example, can be estimated by, $r_{11} = r_{tt}/(t + (1 - t) r_{tt})$, where r_{tt} is the observed reliability for t minutes of observation.

REFERENCES

Adams, J. A. Human tracking behavior. Psychological Bulletin, 1961, 58, 1, 55 - 79.

Alluisi, E. A. Optimum uses of psychobiological sensorimotor and performance measurement strategies. Human Factors, 1975, 17 (4), 309 - 320.

Alluisi, E. A. Methodology in the use of synthetic tasks to assess complex performance. Human Factors, 1976, 9 (4), 375 - 384.

Bachrach, A. J. Psychological research: An introduction. (2nd ed.) New York: Random House. 1965.

Bachrach, A. J. & Kennedy, R. S. Psychological performance testing under water and pressure: Problems and prospects. Bethesda, Md.: U.S. Naval Medical Research Institute, 1977. (In press).

Bradley, J. V. Studies in research methodology. III. The persistence of sequential effects despite extended practice. Wright-Patterson Air Force Base, Ohio, MRL Technical Document 62-60, June 1962.

Broadbent, D. Noise, paced performance, and vigilance tasks. British Journal of Psychology, 1953, 44, 295 - 303.

Cross, K. A. & Bittner, A. C., Jr. Accuracy of altitude, roll angle, and pitch angle judgments as a formation of size of vertical contact analog display. Point Mugu, CA: Naval Missile Center, January 1969 (PM-69-2).

Fleishman, E. A., & Hemple, W. E., Jr. The relation between abilities and improvement with practice in a visual discriminatory task. Journal of Experimental Psychology, 1955, 49, 301-312.

Fregly, A. R., & Kennedy, R. S. Comparative effects of prolonged rotation at 10 rpm on postural equilibrium in vestibular normal and vestibular defective human subject. Aerospace Medicine, 36, 12, 1965.

Guedry, F. E., Jr., Kennedy, R. S., Harris, C. S., & Graybiel, A. Human performance during two weeks in a room rotating at three rpm. Aerospace Medicine, 35, 11, 1964.

Graham, C. H. Visual perception. In S. S. Stevens (Ed.) Handbook of experimental psychology. New York: Wiley & Sons, 1951.

Graybiel, A., Kennedy, R. S., Knoblock, E. C., Guedry, F. E., Jr., Mertz, W., McLeod, M. E., Colehour, J. K., Miller, E. F., II, & Fregly, A. R. Effects of exposure to a rotating environment (10 rpm) on four aviators for a period of twelve days. Aerospace Medicine, 36, 8, 1965.

Grodsky, M. A. The use of full scale mission simulation for the assessment of complex operator performance. Human Factors, 1967, 9 (4), 341 - 348.

Kennedy, R. S. Individual differences in auditory vigilance performance on the band-pass ability (B-PA) test: Some theoretical considerations. Presented at Human Factors Society annual meeting, San Francisco, CA, October 1970.

Kennedy, R. S. A sixty-minute task with 100 scoreable responses. Naval Aerospace Medical Center, Pensacola, Florida, NAMI-1045, 1968.

Kennedy, R. S. A performance assessment in exotic environments: A flexible, economical, and standardized vigilance test. Paper presented at the Fifteenth Annual Human Factors Society Meetings, New York, October 1971.

Kennedy, R. S., & Bruns, R. A. Consideration for the utilization of a flexible economical, vigilance test to assess performance in exotic environments. Presented at the October 1975 Aerospace Medical Panel Specialists' Meeting, Ankara, Turkey.

Kennedy, R. S., & Hutchins, C. W. Relationships between physical fitness, endurance, and success in flight training. Naval Aerospace Medical Center, Pensacola, Florida, NAMI-1088, in press, 1971.

Kennedy, R. S., Moroney, W. F., Bale, R. M., Gregoire, H. G., & Smith, D. C. Motion sickness symptomatology and performance decrements occasioned by hurricane penetrations in C-121, C-130, and P-3 Navy aircraft. Aerospace Medicine, 43, 1235 - 1239, 1972.

Kennedy, R. S., Tolhurst, G. C., & Graybiel, A. The effects of visual deprivation on adaptation to a rotating environment. NSAM-918. NASA Order No. R-93. Pensacola, FL: Naval School of Aviation Medicine, 1965.

Kennedy, R. S. PETER for Mentation Mensuration. A point paper, Naval Aerospace Medical Research Laboratory Detachment, New Orleans, LA, December 1977. (In press).

Parker, J. F., Jr. & Fleishman, E. A. Ability factors and component performance measures as predictors of complex tracking behavior. Psychological Monographs, 1960, 74 (16, Whole No. 503).

Parker, J. F., Jr. & Fleishman, E. A. Use of analytical information concerning tasks requirements to increase effectiveness of skilled training. Journal of Applied Psychology, 1961, 45, 295-303.

Parker, J. F., Jr. Use of an engineering analogy in the development of tests to predict tracking performance. The Matrix Corporation. (Office of Naval Research Contract No. ONR-3065(00)). February 1964.

Reilly, R. E. & Cameron, B. J. An integrated measurement system for the study of human performance in the underwater environment. ONR Contract N0014-67-C-0410, December 1968.

Rose, A. M. Human information processing: An assessment and research battery. University of Michigan, Technical Report No. 46.

Rose, A. M. Human information processing: An Assessment and research battery. Ann Arbor, MI., University of Michigan, Doctoral dissertation, 1974, also published as AFOSR-PR-74-1372 (AD-785-411).

Rosenthal, R. Experimental outcome--orientation and the results of the psychological experimentation. Psychology Bulletin, 1963, 61, (6), 405-442.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. Decision processes in perception. Psychological Review, 1961, 68, 301 - 340.

Teichner, W. H. Quantitative models for predicting human visual/perceptual/motor performance. New Mexico State University/Office of Naval Research - Technical Report 74-3, Las Cruces, New Mexico, October 1974.

Theologus, G. C., Wheaton, G. R., Mirabella, A., Brakler, R. E., & Fleischman, E. A. Development of a standardized battery of performance tests for the assessment of noise stress effects, NASA CR 2149, Washington, D. C., 1973.

Thorndike, R. L. Personnel selection: Tests and measurements techniques. New York: Reilly, 1949

Woodrow, H. The effect of practice on groups of different initial ability. Journal of Educational Psychology, 1938, 29, 268-278(b).

Woodrow, H. The relation between abilities and improvement with practice. Journal of Educational Psychology, 1938, 29, 215-230(a).

Woodworth, R. S. & Schlosberg, H. Experimental psychology. New York: Henry Holt & Co., 1954.

ABOUT THE AUTHOR

Robert S. Kennedy has been an aerospace experimental psychologist since he entered the Navy in 1959. He received an MA in experimental psychology from Fordham University in 1959 and a Ph.D. from the University of Rochester in 1972. His previous military experience includes two tours in the Aerospace Psychology Division at the Naval Aerospace Medical Institute, Pensacola, Florida, where he conducted research on vestibular function, motion sickness, vigilance, and habituation in exotic environments; one tour at the Behavioral Sciences Department at the Naval Medical Research Institute, Bethesda, Maryland, working on the Man-in-the-Sea program; one tour at the Pacific Missile Test Center, Point Mugu, California; and one tour at the Air Development Center, Warminister, Pennsylvania, where he worked mainly on the development, test, and evaluation of airborne weapons systems from the standpoint of human factors engineering. Presently, he is the Officer-in-Charge of the Naval Aerospace Medical Research Laboratory Detachment working on human performance mensuration in unusual environments, specifically ship motion.

DATE
ILME