NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN DIEGO, CALIFORNIA 92152

NPRDC SR 78-8

APRIL 1978

HUMAN FACTORS IN OPERATIONAL SYSTEM TESTING: A MANUAL OF PROCEDURES

Special Report 78-8 April 1978

HUMAN FACTORS IN OPERATIONAL SYSTEM TESTING: A MANUAL OF PROCEDURES

D. Meister

Reviewed by
F. Muckler

Approved by
James J. Regan
Technical Director

Navy Personnel Research and Development Center
San Diego, California 92152

FOREWORD

This manual was developed under Exploratory Development Work Unit Number 521.101.03.10: U. S. Marine Corps Human Factors Analysis. It describes procedures for personnel performance evaluation in the context of an Operational System Test (OST).

Although these procedures have been developed specifically for the U. S. Marine Corps, it is believed that they are applicable to personnel performance evaluation in all military services.

The cooperation of the U. S. Marine Corps Headquarters (MPI-20) and of the Marine Corps Development and Education Command (OT&E), Quantico, Virginia is much appreciated.

Section EIGHT: Introduction to Statistical Methodology was authored by Dr. Jules Borack.

J. J. CLARKIN
Commanding Officer

CONTENTS

SECTION ONE--INTRODUCTION

In accordance with MCO P5000.11, Test and Evaluation of Systems and Equipment for Operating Forces of the Marine Corps, the U. S. Marine Corps performs Operational System Tests to determine that new man-machine systems entering the Marine Corps inventory will satisfy Corps requirements.

Such systems consist of equipment; personnel who will operate and maintain the equipment; and the procedures, logistics, technical documentation, etc. necessary to support the personnel. The evaluation of each new system therefore must include a determination that personnel can in fact efficiently operate and maintain that system in accordance with Marine Corps requirements.

Experience with many systems acquired not only by the Marine Corps but also the other Services indicate that, if personnel have difficulty performing their tasks, the new system, no matter how sophisticated and well engineered, will not function as effectively as it should. Each equipment and system imposes its own burdens on the personnel operating and maintaining that system. To use a very simplified example, if a rifle is excessively heavy, it will take longer for the rifleman to aim the rifle and his rate of fire will thereby be reduced. The more complex the new system, the more likely that it will contain features that increase the difficulty of using it.

Despite this, military testing of new systems often tends to ignore the personnel performance aspect. Many testers feel that the human is so flexible that he can adapt to and overcome equipment characteristics that create difficulties. Ordinarily this is correct, but combat--in which the new system may eventually be tested--is not an ordinary situation and severely stresses the Marine. Under stress conditions, minor equipment deficiencies that could otherwise be lived with become more difficult to overcome. Under these circumstances the human is not as flexible and adaptive as he is ordinarily.

Those authorized to evaluate newly acquired systems have a significant responsibility for uncovering minor equipment deficiencies. They are in a key position to improve the quality of Marine Corps equipment by determining where inadequacies exist and by recommending ways of eliminating them.

The title of this manual is "Human Factors in Operational System Testing: A Manual of Procedures." Human Factors refers to the entire complex of elements that affect personnel performance. These elements include (1) the way in which the equipment has been designed (equipment characteristics), (2) the procedures developed to operate and maintain the equipment, (3) the training provided to enable effective operation and maintenance, (4) the performance aids supplied to assist operators and maintainers (e.g., technical manuals), and (5) the environment in which the equipment will be used. The discipline of Human Factors has, on the basis of research conducted and experience gained over the 35 years since World War II, established personnel standards for each of the above elements.

A major part of Human Factors work is the development of procedures for testing to determine whether new equipment and systems meet these personnel standards. The procedures in this manual have been developed to permit Marine Corps test personnel to evaluate the new system for its human factors adequacy. They have been specially designed to require as little specialized Human Factors knowledge as possible. Where significant personnel-related problems arise, however, a Human Factors specialist should, of course, be consulted.

No one can force the test planner and test conductor to include these procedures in his test operations. If they are ignored, however, the resultant test and the conclusions derived from the test will be inadequate to answer the basic question of whether the system is one that the Marine Corps wants in its inventory. Further, the chances are excellent that the system, if accepted, will have degraded personnel capabilities.

SECTION TWO--THE PERSONNEL PERFORMANCE TEST PLAN

This section provides information on the Personnel Performance Test Plan, including its purpose, a step-by-step description of its various sections and requirements for completing those sections, a model to illustrate the material to be included in such a plan, and a procedure for developing quantitative personnel performance criteria.

Introduction

Operational System Testing (OST) is the process by which a military service verifies that a newly acquired system can in fact perform in accordance with military requirements. OST is required of the U. S. Marine Corps in its Systems Acquisition Management Manual (MCO P5000.10). An essential part of that verification process is the determination that system personnel--those who will operate and maintain the system--can perform required tasks to a required level of capability.

The first and essential step in the measurement of personnel performance is to develop a Personnel Performance Test Plan. The purposes of this section are to describe, in simple, step-by-step fashion, how to develop that test plan and the major points to be covered, and to illustrate it with several examples. This test plan must, of course, satisfy the format established in MCO P5000.11, Test and Evaluation of System and Equipment For Operating Forces of the U. S. Marine Corps.

Development of a Personnel Performance Test Plan is only the first of the activities required in measuring personnel performance during OST. It is, however, critical to the planning of a number of additional activities, including the following:

1. Development of performance criteria.

2. Selection of performance measures.

3. Selection of subjects (the system operators and maintainers whose performance is to be measured).

4. Determination of methods of measurement.

5. Methods of analyzing and reporting performance data.

The test plan is the basic document describing how the personnel part of the system will be tested.

The Purpose of Personnel Performance Testing

Before a new system can be accepted into inventory by a military service, it is necessary to verify that the system can perform its mission to specified requirements. Verification is a short-hand phrase which includes the following test goals:

1. To assess the accomplishment of system development objectives.

2. To ensure that systems, equipment, and personnel meet established requirements.

3. To forecast how the system will perform in an actual operational situation.

4. To ensure the effective integration of all elements of the system (including equipment, personnel, technical data, supplies, etc.).

5. To detect operational and engineering deficiencies in time for changes to be incorporated before significant production build-up.

6. To provide data and operational analyses for application to current and future systems.

7. To identify manpower and personnel resources needed to support the operational system.

8. To provide information to validate forthcoming training programs.

All of the above involve personnel testing. OST assumes that:

1. The system being evaluated is a prototype of the operational system and is configured as much as possible as it will be in its eventual operational use. If the system being tested deviates significantly from its operational configuration, the conclusions derived from testing will not describe what the system will be able to do operationally. What this means in effect is that personnel operating and maintaining the equipment during OST should be required to perform as much as possible as they would once the system is put into actual use. In other words: Don't take short cuts that would not be allowed in actual operation; don't give test personnel assistance they would not ordinarily have; and carry the test out to its ultimate conclusion and don't abort because something goes wrong (unless it imperils the system and its personnel). If the system is not operated according the operational procedures, conclusions cannot be drawn about how that system will function in the real world. Of course, it is impossible to fully simulate a combat situation, but test conditions should approximate this as much as possible.

2. Personnel operating and maintaining the system being evaluated should have characteristics similar to those that operational personnel will have and are required to perform as they would operationally. If the test conductor has a choice, he should select test personnel who are "average" in their past performance, not those whom he considers to be most proficient. Test personnel who, as a group, have a range of capability are most desirable. The most proficient personnel will make the system appear to be more efficient than it will actually turn out to be when it must be operated by average Marines. If there are any weak points in the system, it will be easier to find these with average personnel than with those who are much more proficient; the latter can compensate for these weak points, the former cannot.

3. All aspects of the system (i.e., equipment, personnel, logistics, procedures, and data) should be measured as an integral part of the evaluation. How the system functions is determined by all its elements. Ignoring one or more of these elements means that the system will not function realistically. Moreover, one cannot extract a system element and test it apart from other elements.

4. Once OST has begun, let the test personnel assume control of system operations; that is, the system should be allowed to perform without interference from Test Management (except to avoid conditions that may affect personnel safety or the integrity of the system).

The term "system" as used henceforth in this section includes not only an entire system, but any subsystem or equipment unit of the system (together with its operating personnel).

Logic requires that OST involve measurement of personnel performance. Every system procured for a military service includes not only equipment, but also operating and maintaining personnel, together with the procedures for these operations. Personnel form a subsystem of the total system; and the total system cannot perform its mission adequately unless all its subsystems function in accordance with requirements. The system will therefore not function properly unless its personnel also function properly. System adequacy cannot therefore be verified unless it is certain that system personnel can operate and maintain the system to specified standards.

Tests of engineering capability designed to measure physical characteristics of the system (e.g., fuel consumption, rated engine power, missile penetrating force) do not automatically measure personnel performance. Consequently the OST must make special provisions for measurement of personnel performance. Unless this is done, the OST will lack validity. The overestimation of system capability by developers often results from failure to factor in the ability of personnel to utilize the system. Since personnel performance often degrades system effectiveness, lack of personnel performance measurement data requires the system evaluator to assume perfect performance on the part of personnel--which obviously is not often true.

It is often assumed by those unfamiliar with the process that personnel performance measurement is an expense added to the already burdensome cost of OST. However, since personnel performance is inherent in system operations, and since the system must be exercised as it would be in routine operations, personnel performance measurement can be coordinated with other on-going aspects of the OST. It usually requires no additional test time beyond that required for on-going, already scheduled tests. There may be occasions when a special question must be examined and the overall OST does not contain provisions for examining the question; a special test may then be required. The number of such special tests will, however, be minimal.

In most cases, personnel performance measurement also requires no, or only minor, special instrumentation. It does, however, require the assignment of a limited number of personnel to collect and analyze the necessary data. The cost of this personnel time is relatively slight considering the importance of securing a quantitative measure of personnel effectiveness.

The measurement process does, however, require special test procedures that are described in the Personnel Performance Test Plan. This plan is not the overall system test plan but merely one part of it; most often it will be incorporated in the test plan as a special annex.

Development of the test plan is not merely a pencil-and-paper exercise. It is needed for the following reasons. First, writing the plan ensures that those who are responsible for performance measurement and others involved in the process (like data collectors) know exactly what they are supposed to do. Failure to write the plan often means that essential measurement elements are overlooked. Second, Test Management personnel often require that such a plan be written as a preliminary to conducting the OST. This is necessary if they are to know what is going on.

Outline of Personnel Performance Test Plan

Figure 2-1 provides an abbreviated outline of a Personnel Performance Test Plan (a full-scale test plan would contain much more detail). The examples provided under the various sections of the plan do not refer to any existing U. S. Marine system or equipment. They may imply a rather complete field test, which would be appropriate for the evaluation of a major system like a tank, an aircraft, or a fire direction center. The size and complexity of the system being evaluated (including the functions to be performed by personnel) will usually determine the scope of the testing. If a subsystem or a unit of a system is under evaluation (e.g., a single console), it may not be necessary to exercise the entire system of which the item under test is only an element. For example, if one were evaluating a new rifle, it would not be necessary to engage in regimental maneuvers to test that rifle. The scale of testing is determined by the number and type of inputs needed to reproduce the operational situation in which the system, subsystem, or unit will be used. In the case of the rifle, for example, this would probably be only a firing range reproducing the types of targets and the terrain in which one would use the rifle.

The same qualification applies when the purpose of the test--the specific question to be answered--is highly specialized. Assume that the evaluator wished to check to make sure that the interior environment of a new tank--temperature, humidity, air flow--met the standards of MIS STD 1472B (Department of Defense, 1974),[1] which is the governing document for the human engineering characteristics of man-machine systems. If that were the only question to be answered by the test, it might not be necessary to exercise the tank in simulated combat in order to secure an answer. Driving the tank for specified periods of time during the day and night might provide the needed data. As in the case of a unit of a system, the scale of testing required to answer a particular question is determined by the number and type of inputs needed to provide a reasonable simulation of the operational situation.

---

[1]Department of Defense. MIL STD 1472B. Military Standard, Human Engineering Design Criteria for Military Systems, Equipment, and Facilities, 31 December 1974.

1.0 PURPOSE

1.1 General. Example: Verify that system personnel can perform required tasks.

1.2 Specific. Example: Determine the type and magnitude of errors made by personnel; determine the effect of low temperature (Arctic) conditions on personnel ability to maintain a tank.

2.0 DESCRIPTION OF SYSTEM BEING EVALUATED

2.1 List of equipments to be operated (maintained) by personnel and for which personnel performance data are to be collected.

Example: _____ gun mount; _____ grenade launcher.

2.2 List of equipment tests during which personnel performance data are to be collected.

Example: Installation and checkout of the _____ fire direction console.

2.3 List of tasks for which personnel performance data will be collected.

Example: Alignment of the _____ theodolite.

2.4 Applicable Technical Manuals or other procedures.

Example: TM _____, Operation of the _____ machine gun;

TM _____, Operation of the _____ landing vehicle, personnel.

3.0 SPECIAL COMPARISONS (optional)

Example: Comparison of daytime and nighttime reconnaissance in jungle terrain.

4.0 CRITERIA AND MEASURES

4.1 Personnel performance criteria. Example: Receive, code, and transmit between 12 and 15 messages per hour; fire minimum of 5 rounds within 7 minutes.

4.2 Personnel performance measures. Example: Time taken to load hand-held missile; officer evaluation of squad reconnaissance performance.

Figure 2-1. Outline of the Personnel Performance Test Plan.

5.0 DATA COLLECTION METHODS

5.1 Data collectors

5.1.1 Number. Example: 4

5.1.2 Tasks to be performed. Example: Record start/stop time in operation of laser tracking set.

5.1.3 Training (if required). Example: All data collectors will receive 3 hours instruction in gathering data on the retractable machine gun (see training schedule appended).

5.2 Data collection forms. Example: Data sheet for retractable machine gun (appended); Post mission debriefing questionnaire for tank driver (appended).

5.3 Data collection procedures. Example: See Appendix A.

5.4 Instrumentation (only if required). Example: 2 tape recorders with 3 rolls of tape per data collector.

6.0 SUBJECTS

6.1 Number. Example: 3 squads.

6.2 Required characteristics. Example: ALL subjects will have 20/20 vision (corrected) and will have been qualified in operation of the retractable machine gun.

7.0 DATA ANALYSIS

Example: Determine mean number (and standard deviation) of messages transmitted between forward observers and batteries; develop equation relating gun loading speed and operator errors.

8.0 TESTING SCHEDULE

Example: Concurrent with other tests.

Figure 2-1 (Continued).

The principles described in this section apply regardless of the size of the unit being tested or the scope of the question being asked. In all cases, it is necessary to develop a Personnel Performance Test Plan, including the analyses involved in developing that plan. If the size of the unit to be tested is small, or the questions to be answered are few, the scope of the test plan can be scaled down. However, all sections of the test plan must be completed.

The following paragraphs describe the various sections of the test plan. The reader should refer to Figure 2-1 as each section is described.

Section 1.0--PURPOSE

This section describes the purpose of the Personnel Performance Test. The test has two purposes, general and specific. Of the two, the latter is by far the more important.

Although the purpose of personnel performance measurement may seem obvious at first glance, it is not. Thus, that purpose must be specified in detail.

The general purpose is to verify that personnel can perform their assigned tasks within the new system to meet system requirements and to ensure that personnel functions are effectively integrated with other system elements (see preceding section entitled "Purpose of Personnel Performance Testing"). The general purpose points to the necessity for specifying (preferably in quantitative form) the standards that the system and the personnel must meet. It emphasizes the necessity of comparing actual and desired personnel performance.

Unfortunately, the general purpose does not indicate what and how one should measure. Although it is important to emphasize that the general purpose is to verify personnel performance adequacy, this purpose is not specific enough to be very useful.

What and how one should measure can be determined only by breaking down the general purpose into specifics. A specific purpose of personnel performance measurement is simply a question about that performance which the test planner wishes to answer. In developing specific test purposes, the planner should list every question he wishes the personnel performance data to answer. Each such question becomes a test purpose. Consequently, personnel performance measurement usually has more than one specific purpose.

If there is a possibility that personnel performance will affect other system elements (the equipment configuration, its procedures, environment, or logistics), then a question must be answered. For example, if it is possible that heat/humidity within a tank may degrade tank personnel performance, then the planner wants to accept or reject this hypothesis. Therefore, he must arrange to measure personnel performance in relation to heat and humidity. A specific purpose of the test plan might therefore be phrased as follows:

Example: To determine the capability of tank personnel to operate tank controls while "buttoned up" under high heat/humidity conditions.

This specific purpose tells the evaluator that he must arrange to collect data on operation of tank controls under high heat/humidity conditions. This means that he may have to take the following action:

1. Install a thermometer inside the tank to determine internal temperature/humidity.

2. Measure operator performance as a function of temperature/humidity conditions.

3. Compare that performance with heat/humidity standards specified in MIL-STD 1472B (DoD, 1974).

None of this would have been implied by the general purpose. It is therefore important to list all the specific purposes for which personnel performance measurements will be made.

Section 2.0--DESCRIPTION OF SYSTEM BEING EVALUATED

Describing the system to be evaluated may appear unnecessary to those who are familiar with it. However, in most systems of any size, only some of the system equipment and only some of the tasks involved in operating that equipment will be of interest for personnel performance measurement. These should be identified and described.

If those system operations for which personnel performance data will be collected are not described, investigators may not have a clear idea of how to accomplish the measurement. The larger the system, the more necessary this section is.

It is not necessary to provide highly detailed equipment/task descriptions. If Technical Manuals (TMs) are available at the time of testing, they will provide this detail. If TMs are not available, other system design documents describing relevant equipment should be listed. In any event, it is necessary merely to list the equipment and tasks.

List all equipments for which personnel performance data are to be collected, even if this is only one equipment. For these data refer to TMs and operating procedures. Later on, in describing the criteria and measures the test planner will use (section 4.0), he will consider the characteristics of the equipment listed, because this will dictate his data collection measures.

Example: If the equipment is a radar set, the evaluator may wish to measure number of targets detected or gain level used on the CRT; if the equipment is an artillery piece, he may wish to measure number of rounds fired, speed of loading, etc.

Personnel performance data are collected as part of operational exercises or equipment tests. Ordinarily one would expect to collect such data whenever an operational exercise is conducted but, if this is not possible, those exercises during which the evaluator intends to collect data should be listed. This applies even more so to equipment tests that are being planned, because some equipment tests may not provide a proper environment for personnel performance. A test in which only a vehicle's engines are run up would hardly provide an adequate opportunity to collect personnel data. Because, in almost all cases, personnel performance testing will be "piggybacking" data collection on other tests, the planner should determine which ones he will use and list these. For these data, he should refer to the master test schedule. Whenever the evaluator finds an operation of interest for personnel performance measurement, he should require data collectors to observe the operation and collect relevant data.

Example: If the overall system test plan contains a number of tests which involve significant amounts of personnel interaction, these should be pinpointed as ones for which personnel performance data should be collected.

Not every task performed by operators in a particular test may be of concern to the evaluator. All operations involving operator/maintainer actions should ideally be the subject of personnel performance data collection. If the number of data collectors available is insufficient, however, it may be impossible to record data from all tasks. Under these circumstances, tasks that have the following characteristics should be singled out and identified so that available data collectors can be assigned to them:

1. Tasks that, in the opinion of system engineers or experienced operational personnel, are especially important for satisfactory completion of the system mission.

2. Tasks in which human error or other personnel inadequacies could have a significantly negative effect on the system output.

TMs or other procedural documents describing the tasks for which personnel performance data will be collected should be identified. Hopefully, these documents will contain the personnel performance standards to be used in development of criteria and measures (for use in section 4.0). At the very least, data collectors should be aware of these documents as reference sources and should familiarize themselves with the documents in preparation for data collection.

Section 3.0--SPECIAL COMPARISONS

This section describes Special Comparisons to be considered in performing a personnel performance evaluation.

The basic idea of personnel performance measurement in an operational system test is to determine whether operators can perform their tasks in accordance with system requirements. There may, however, be operational

conditions inherent in the system mission that affect how personnel will perform. The planner may therefore wish to perform special analyses to determine how these conditions affect personnel performance.

Note that these conditions are inherent in system operations and therefore do not require the setting up of special tests to collect required data. For information on such conditions, refer to TMs or other documents describing system operations.

Special comparisons usually imply a comparison of performance under alternative conditions.

> Example: Differences in operator performance between daytime and nighttime operations; between different work shifts; between different EM ranks (when these perform the same tasks).

After such conditions are identified, all one need do is to assign data collectors to observe performance under these conditions and to record appropriate data. If data are to be collected under all mission conditions, comparison data will be collected automatically. However, in writing the Data Analysis section of the Test Plan (section 7.0), the planner should note that a statistical comparison will be made between the conditions of interest. Failure to make such a comparison will reduce the value of the personnel performance measurement. If such comparison conditions do not exist, this section of the test plan should be noted as Not Applicable (N/A). It is assumed that the observations will be made unobtrusively and will therefore not bias the comparison.

The usefulness of such a comparison is that, if operator performance is in fact significantly less effective under one or more of the conditions being compared, analysis of the reasons for this may suggest ways of improving operator performance under those conditions.

Section 4.0--CRITERIA AND MEASURES

Criteria are standards to be met. Two types must be considered: those that describe the performance that personnel are expected to achieve; and those that reflect the adequacy of data collection. The first is more important than the second.

Personnel performance criteria indicate what operators must do in order to satisfy overall system requirements. This information can sometimes be secured from a statement of the system mission or, at a more detailed task level, from operating and maintenance procedures described in TMs or similar documents.

> Example: Operator is expected to record and decode 16 messages per hour; to detect all targets at a distance of 1000 meters.

Criteria are important because they serve as performance standards against which actual operator performance can be compared. This comparison

enables the evaluator to say whether or not the performance of system personnel is satisfactory. If quantitative criteria are lacking, it is impossible to state unequivocally that personnel performance meets system requirements. (The previous sentence does not ignore the usefulness of qualitative data in supplementing the quantitative data. As is explained in Section Three of this report, qualitative data are necessary to help explain the significance of conclusions based on quantitative data; and occasionally no quantitative data can be secured. However, qualitative data alone are insufficient for evaluating personnel.)

Unfortunately, many statements of system requirements do not include quantitative requirements for personnel. Qualitative criteria (e.g., "riflemen will perform reconnaissance missions as quickly as possible") are almost useless as standards, because one cannot compare quantitative data (e.g., mean mission duration was 23.5 minutes) with qualitative criteria. If quantitative criteria are not explicitly specified, it will be necessary for the planner to derive them by using the judgment of experienced military personnel, engineers, or others who are familiar with system requirements (see Development of Quantitative Performance Criteria, p. 2-28. Whatever the source, personnel performance criteria must be described. If there are many detailed performance standards, these can be listed in an appendix to the Personnel Performance Test Plan. At the very least, TMs or other procedures in which the personnel standard can be found must be listed.

Criteria of data collection describe how much data should be gathered. One can collect too little or too much data. The first situation is more serious than the second. With too little data, it is impossible to come to any valid conclusions about performance. Gathering too much data is inefficient, but at least one has the minimum data on which valid conclusions can be based. This is not meant to imply that the more data one has, the more valid one's conclusions.

In the context of the Operational System Test, it is far more likely that too little data will be collected. There are many reasons why this is likely. The time allotted for testing is often too short to permit operators to repeat tasks more than once; equipment may break down and further curtail measurements; emergencies may arise that divert test time to other activities. Under those circumstances, the test director must take what he can get.

However, one principle can be asserted. At least two data items (one repetition of the original event) are required for every operator task for which data are being collected. The size of the data sample needed can be determined by statistical means (see Section Seven of this report). But, in most cases, data collection opportunities will not satisfy that requirement.

The measures (i.e., the types of data) to be recorded should also be specified in the test plan. This will permit the test director to cross-check these measures against the tasks (section 2.3) that require these measures. Failure to list specific measures will often result in failure to record required data.

Measures are derived from two sources: specific test purposes and performance criteria. Test purposes and performance criteria require certain data; and these data call for specific measures. Some criteria and the measures they require are, however, only implied by the system mission. The test director may find it necessary to analyze the logic of system operations in order to extract these criteria and measures. It is also possible to develop measures in the absence of criteria, but such measures may not bear any meaningful relationship to system operations.

Example: During a test, it is possible to record all internal and external communications within a system. In the absence of criteria that specify that personnel should respond to so many incoming calls and make so many outgoing calls within a specified time period, all one can say after analyzing these communications is that personnel did in fact respond to and make so many calls.

Almost all personnel performance measures fall into three general categories: (1) accuracy (or the reverse, errors), (2) duration of responses, and (3) reaction time to initiating events. However, if these are to be useful, they must be specified in terms of the task to be performed and the equipment being operated.

Example: Accuracy (distance from the target in firing the _____ semiautomatic rifle); duration (time taken in loading an artillery piece).

It is also desirable to indicate when the measure will be recorded, if this information is not implicit in the description of the measure.

Measures may vary in their detail, depending on the task level they describe. In firing a rifle, for example, one could record the action of the finger in squeezing the trigger; this would be the most detailed level of measurement. Or one could record the rifleman's error in attempting to hit the target; this would be a higher-order and therefore a grosser measurement level. If a team is involved in the performance one is measuring, the team measure (e.g., whether or not the squad accomplished its objective) would be at still a higher level than measurement of the individual rifleman's performance. The question the evaluator wishes to answer should determine the level at which he will measure. He may wish to measure at several levels simultaneously, but he should measure only when there is a meaningful evaluation question to be answered. Obviously, if he wishes to evaluate squad performance, he would not measure the pressure with which the individual rifleman squeezes the trigger.

The level at which one measures may determine how the data are collected. For example, instrumentation would probably be required if one wished to determine squeeze pressure on the trigger, but not necessarily if one measured error in hitting a target.

Level of measurement is determined by the immediacy of its relationship to the system output. If, for example, the system output one is concerned with is the amount of fire one is placing on a target, then highly

detailed measures like pressure on the trigger would be irrelevant, since squeeze pressure is only indirectly related to amount of fire; rounds expended would be more directly related.

Section 5.0--DATA COLLECTION METHODS

This section deals with who will collect data (e.g., Marines selected from an infantry Battalion; or contractor personnel) and how they will collect it.

Note as a matter of record the number of personnel who are serving as data collectors (section 5.1.1). Where more than one data collector is required, how they are to be scheduled becomes important.

Indicate the task which the data collectors will perform (section 5.1.2). These tasks are not the same as system operating tasks for which personnel performance data are to be collected. Rather, they are the activities involved in gathering information about the task performance being measured.

> Example: At the conclusion of each tank driving cycle, data collectors will administer a questionnaire concerning ride quality.

The level at which these data collection tasks are described need be general only. In the event that there are many data collectors or they have many data collection tasks, section 5.1.2 can be an appendix to the test plan. If several data collectors perform the same data collection task, a single description will suffice.

The reason for describing data collection tasks is to ensure that all personnel know what their responsibilities are. Failure to inform data collectors of their responsibilities will result in lost or erroneous data.

If data collectors must receive special training to enable them to perform their duties effectively, that training should be described (section 5.1.3). In general, all data collectors should receive training in methods of recording data, even if they are specialists in the system being tested. If data collectors are not familiar with the system under test, they will require special training to give them this familiarity, since a data collector cannot function effectively without that information. Data collection training should be oriented toward enabling collectors to recognize the events they have to record and how these should be recorded. At the very least, several data collection "dry runs" with recording forms should be held to habituate data collectors to their tasks.

All data recording forms should be noted in section 5.2. The actual forms themselves should be appended to the test plan. The reason for listing the forms is to remind the test director that these forms must be developed (if they are not already available). Appending the forms to the test plan permits the test director or anyone else to examine them to determine if they are satisfactory for their purposes.

Most data recording tools are of the following types: time and events recording sheets, questionnaires, rating forms, and checklists. Instrumentation used to record data is described in a separate section (5.4).

Unless the data collection procedure is very simple, it should be described in some detail (section 5.3). This description is to let data collectors and all others involved in the evaluation know what is required of them. It should include the following (as relevant) and any other significant aspects of the work:

1. The hours data collectors will work, or the sequence of operations (their beginning/completion) that will determine the data collection period.

2. How data collectors should process their data. For example, do they pass the data on to the evaluation personnel immediately or hold on to it? Do they do any analysis of the data during the performance being evaluated?

3. What data collectors should do if an emergency occurs (e.g., if an exercise is suddenly stopped before it is completed), or if something not covered by operating procedures occurs.

4. The level of detail to which they record data (principally relevant when reporting qualitative observations).

5. The extent to which the data collector is permitted to interact with the personnel whose performance is being evaluated (e.g., the distance they must remain away from participants in the operation).

6. Any equipment data collectors will be required to operate.

Although there are occasions when elaborate instrumentation will be required for personnel performance measurement, in general, instrumentation demands should be minimized, particularly if testing is being conducted in a field environment. It is difficult to operate elaborate instrumentation effectively in that environment, especially when the equipment requires highly controlled conditions.

Data collection instrumentation that is sometimes used in field operations include small magnetic tape recorders for recording communications and hand-held video tape recorders or motion picture video tape cameras for recording events visually. To assess the environmental conditions under which performance occurs, light meters and accelerometers (for vibration effects) may be employed. Sound level meters may be used to record noise levels. Such instrumentation should be used only if data on environmental conditions are necessary and have not been secured in earlier developmental tests. If evaluations are being held in fixed locations, either on the ground or aboard ship (i.e., where greater control can can be exercised over environmental conditions), more elaborate automatic data recording apparatus may be used.

Section 6.0--SUBJECTS

This section describes the subjects whose performance is being evaluated; that is, those personnel who are selected to operate and maintain the system during its testing (also referred to as test personnel).

If the operational system test is to be valid, it must be performed with personnel who are representative of those who will eventually operate and maintain the system. If, for example, subjects are much more or less trained or experienced than eventual users of the system, evaluation results will fail to describe correctly the performance of these eventual users. If the evaluation is performed under military auspices, selection of the appropriate subjects should pose less difficulty than it would if it is performed with contractor subjects. Even under military conditions, however, the evaluator must exercise care in the selection of subject personnel. At the very least, he must examine any special requirements that user personnel will have, and match his subject characteristics to those requirements.

Test personnel requirements include the following:

1. Physical--e.g., vision (20/20), hearing, height, weight, and strength.

2. Aptitude--general intelligence; special aptitudes (e.g., mechanical).

3. Training--e.g., graduation from a required training course.

4. Experience--number of years in a given military speciality.

Attitudinal and motivational factors are important also but cannot be precisely specified. If military personnel are used as subjects, they are, of course, subject to orders; but, even in this case, it is important to explain to personnel why they are being asked to participate in the OST and the importance of their performance.

Most frequently, subject requirements will involve training and experience. For example, it is obvious that, to evaluate a prototype tank, the subjects must be qualified tank drivers and, if driving the new tank requires new skills, they must have received factory training in tank driving. Although the personnel requirements in the example are relatively obvious, others may not be. In such cases, the evaluator may have to determine them from analyzing the personnel requirements information in documents describing the system.

Section 6.1 notes the number of subjects or the organizational units in which they function (e.g., the squad, the platoon). This item is for information only.

Section 6.2 describes required subject characteristics. These are characteristics that, in the evaluator's judgment, will significantly affect the accuracy of the data if not possessed by personnel acting as subjects.

Section 7.0--DATA ANALYSIS

It is essential that the test plan include a procedure for statistical analysis of the data. If the test director waits until after data are collected before developing his statistical plans, the chances are excellent that too much or too little data will be collected and--much more serious--much of his data will be unusable because it will not fit statistical requirements.

All data analysis in system testing involves one or both of two types of comparison. The first compares personnel performance in the system under evaluation with a system requirement; this comparison verifies that personnel can perform required tasks (general purpose, section 1.1). The second comparison is between any special conditions that are tested (e.g., daytime vs. nighttime performance) (see section 3.0). In both cases, the evaluator needs a statistical plan that describes the kind of statistics he will use in relation to which data measures in order to answer which questions.

Because planning an appropriate statistical data analysis is a highly specialized technique, it is advisable for the test planner to allow a statistical expert to prepare this section of the test plan (unless, of course, he is also a statistician). At least the planner should permit an expert to review the statistical design of the study and the entire test plan before it is finalized.

Section 8.0--TESTING SCHEDULE

This section describes the personnel performance data collection schedule. If the system test is very complex, and only some of the system operations will be used as occasions for the gathering of personnel performance data, then a daily, weekly, or monthly schedule of data collection activities should be appended to the test plan.

Example: In one flight navigation testing program for helicopter pilots, pilots of varying levels of experience were to be tested over a year's time. Since pilot navigation performance was hypothesized to depend in part on the appearance of the terrain, it was necessary to systematically arrange the subject schedule so that pilots with different experience levels could be tested during both summer (heavy foliage) and winter (bare trees, snow) conditions.

If the personnel performance data are to be collected during all test events, a detailed subject schedule is unnecessary because the overall test schedule will determine where data will be collected.

Example of Personnel Performance Test Plan--Personnel Performance Measurement in the XM-47 Amphibious Troop Transport and Tank Destroyer

The example of a Personnel Performance Test Plan described in the following paragraphs has been written to illustrate the preceding instructions and to serve as a model for such a test plan. It is based on an

imaginary man-machine system, the XM-47 amphibious tank/transport, which has some of the characteristics of the Marine Corp's LVTP-7 but which does not pretend to any realism in terms of physical engineering details or operational capabilities. The reader should consider this example as representing a type of system the Marine Corps might have in the future (e.g., 1985).

The XM-47 is an amphibious troop transport that also performs land reconnaissance and functions as a tank-destroyer. It has a crew of three: a driver, a missile loader and mechanic, and a gunner. It can carry 10 fully loaded troops or an equivalent amount of supplies from a "beyond the horizon" mother ship to shore, disembark these troops or supplies, and then perform ground reconnaissance and/or antitank duties. It is armed with the following weapons: (1) Typhoon heat-seeking surface to air (SAM) missile for protection against aircraft (range up to 10,000 feet), (2) the Hurricane antitank guided missile (range up to 2 miles), and (3) two 50-calibre machine guns. It has a top speed of 20 knots in water, 50 mph on well-surfaced roads, and 30 mph over sand and fields.

Testing of this system will be performed as part of an amphibious exercise involving adversary units, including troops, artillery, and aircraft.

The material to be included in each of the various sections of the test plan is described below. The author's comments are presented in italics and indented form. These comments would not ordinarily be found as part of the test plan.

Section 1.0--PURPOSE

1.1 General. The purpose of this test is to verify that system personnel can perform all required tasks to accomplish mission objectives.

1.2 Specific.

In any highly advanced system like the XM-47, many questions beyond the general purpose need to be answered. For example, although considerable analysis went into the specification of a crew of three for this vehicle, it is possible that because of workload considerations more personnel will be needed; hence, only realistic simulated-combat exercises will verify that a crew of three is adequate. Similarly, although noise vibration data were gathered as part of developmental testing, these data did not include measurement of personnel effects resulting from the tank environment; hence, the effect of environmental conditions on performance effectiveness remains to be tested. The adequacy of operating and maintenance procedures can be tested adequately only under operational conditions; hence, this question is also pertinent. To repeat what was said previously: every question which is desirable to answer during this test becomes in effect a specific purpose of the test. (One assumes, of course, that such questions are not frivolous.)

The following are the specific purposes for which the personnel performance evaluation will be conducted:

1. To determine whether crew size is adequate to perform all required tasks without overloading personnel excessively.

The qualification "without overloading personnel excessively" is necessary because in many cases personnel can adapt to stressful or undesirable situations but at the cost of increased likelihood of error. Although such stressful tasks may be accomplished, one may not wish to accept the added risk of error.

At the same time, note that there is a certain ambiguity about the qualification "excessively." It will be necessary for the test planner to determine and specify a criterion of acceptable performance (e.g., number of errors) under overload conditions.

2. To determine whether environmental conditions (heat, noise, lighting, vibration) within the XM-47 will permit operators to perform all required tasks without excessive overload.

3. To determine whether the environmental conditions noted in (2) above will permit troops when transported to shore to perform necessary combat functions without loss of efficiency.

4. To determine that XM-47 missile weapons can be fired adequately.

This implies all the personnel functions involved in successful operation of XM-47 missiles: detection, identification, and hit probability. Again, it is assumed that quantitative criteria for these functions exist or can be determined.

5. To determine that operating procedures can be performed correctly and adequately.

Criteria for these should be specified in section 4.1.

6. To determine the frequency and types of operator errors made and their causes.

7. To determine that the driving characteristics and "rideability" of the XM-47 are adequate.

8. To determine the effects of any fatigue occurring during a routine mission.

9. To determine the adequacy of seat design in the XM-47.

The above list of purposes is not necessarily complete; it is only illustrative. Moreover, these specific purposes will require particular measures and measurement procedures.

The above purposes can, if desired, be broken down even further. For example, many elements make up adequate task performance, such as reaction time to enemy attack or a maximum duration for performing a specific task. Should these elements be pulled out as specific purposes of the test? Not necessarily, as long as the measures developed to satisfy these purposes are detailed enough to provide desired data.

Section 2.0--DESCRIPTION OF SYSTEM UNDER EVALUATION

A general description of the system (e.g., the previous paragraphs describing the XM-47) need not be provided here, since it is reasonable to assume that all concerned are familiar with the general characteristics of the system. However, the specific equipments and tasks with which the personnel performance evaluation is concerned would be listed.

2.1 Equipments Operated (Maintained) by Personnel. Personnel performance data will be collected with regard to the following equipments:

a. Controls and displays for driving the XM-47 (see TM listing below for equipment details).

b. Optical displays for locating ground targets and for control during firing of the Hurricane antitank missile.

c. Controls for loading and firing the Hurricane antitank missile.

d. Radar display for detection of aircraft targets.

e. Controls for firing Typhoon surface to air missile.

f. Radio set 4FU.

g. .50-calibre machine gun.

h. Operator seats and restraint harnesses.

Evaluations will be performed with regard to the following questions: Are the seats comfortable for prolonged missions; and do they provide sufficient arm/leg room to operate controls?

i. Internal XM-47 environment.

The definition of equipment operated by personnel extends to the working environment. It requires consideration of the work space available to operators, controls for adjusting lighting and temperature, etc.

Implications of the items in this section for the evaluation include use of human engineering checklists (see section four) to evaluate arrangement of controls and displays; and personnel responses (in the form of questionnaires and rating scales) concerning ease/difficulty of operating controls and the driving qualities of the XM-47.

2.2 Equipment Tests During Which Personnel Performance Data Will Be Collected. Personnel performance data will be collected during amphibious/ground attack exercises (Blue Force) conducted the week of March 2nd, 1985 at Ocean Peninsula. Data will be gathered on all XM-47 operating cycles. Two prototype vehicles will be utilized, alternating on missions. XM-47 will be subject to control of CO, Blue Force. Two XM-47 crews comprised of three men each will be used in alternation. Start of an XM-47 cycle will be positioning of the vehicle in the well of an LSD. Conclusion of the cycle will be the recall signal as received over XM-47 radio. Approximate duration of the XM-47 mission: 3 hours. Total number of XM-47 mission cycles: approximately eight (two missions per 24-hour interval). On half of these missions, the XM-47 will be loaded with 10 fully equipped troops; on the other half, with supplies. Loading and off-loading of the troops/supplies will be included in mission functions being evaluated. Upon command, XM-47 will drive from the LSD to shore, disembark the troops or supplies, and will conduct reconnaissance/antitank operations against an enemy force of infantry, artillery, tanks, and aircraft. Half the XM-47 operations will be conducted at daytime, the other half at night.

This description has certain implications for the design of the Human Factors evaluation. The fact that the evaluation will take place during simulated combat operations means that certain unprogrammed events (such as enemy air attack) will occur which the data collector must record. Since two vehicles and two different crews will be used in alternation, a statistical comparison of performance between the two vehicles and crews must be conducted. If, for example, statistically significant differences between the performance of the two crews are found (which is unlikely but might occur if the two crews were not given equivalent training), then data cannot be combined and an analysis must be performed to determine why the differences occurred. Since the missions will be divided between troop carrying and supply transport, a comparison of performance between these two types of missions is required, although on the surface it would not seem as if

this factor alone would be important. The fact the XM-47 operations will be divided between daytime and nighttime also requires a statistical analysis to determine if performance differs significantly between these two conditions, and if so, why. Since different geographical terrain conditions are involved (sea, shore, sand, hard surface road, and field), comparisons of performance under these conditions must be performed. Since approximately eight missions will be conducted, mission-to-mission performance should be analyzed. Questions addressed to test personnel will have to be developed in the light of each of the conditions entering into the evaluation.

2.3 Tasks For Which Personnel Performance Data Will Be Collected. Data will be collected concerning performance of the following tasks:

a. Premission checkout.
b. Driving the XM-47.
c. Loading/unloading troops/supplies.
d. Communications over radio and intercom.
e. Operation of the .50-calibre machine gun.
f. Detection of ground and air targets.
g. Firing at ground and air targets.
h. Any required corrective maintenance during the mission.
i. Postmission checkout.

This list ties the evaluation to specific personnel operations. For example, pre- and postmission checkouts are mentioned here for the first time. If such a list were not developed, it is possible that important tasks would be overlooked by the data collector. This list can also be compared against the data recorded and thus serves as a check on the completeness of those data.

2.4 Applicable TMs and Other Documents. The following list of TMs applies to the equipments and tasks being evaluated:

a. TM _____: Checkout Procedures for the XM-47.
b. TM _____: Operator's Manual for the XM-47.
c. TM _____: Operation of the Typhoon surface-to-air missile.
d. TM _____: Operation of the Hurricane Antitank Missile.
e. TM _____: Operation of the 4FU radio set.
f. TM _____: Operation of the .50-calibre Machine Gun.

Other applicable documents include:

a. Rpt. _____: Mission Requirements of the XM-47 Transporter/Antitank Vehicle.
b. Rpt. _____: Function Analysis of XM-47 System Operations.

This list tells data collectors which documents they should refer to in order to secure information about these individual weapons and the features of equipment operation tasks which they should record. It can serve also as a source of information concerning criteria and measures (see section 4.0). Data analysts will probably find these documents useful during the data analysis and writing of the final test report. All data collectors should be issued with a copy of applicable TMs.

Section 3.0--SPECIAL COMPARISONS

The following major comparisons will be made:

a. Performance under daytime vs. nighttime conditions.
b. Between vehicles.
c. Between crew 1 and crew 2.
d. Between troop carrying and supply missions.
e. Between successive operations (mission 1 vs. 2 vs. 3, etc.).
f. Driving qualities on water vs. sand vs. surfaced road vs. field.

This list clarifies the specific analytic comparisons to be made. Previously they had merely been implied by section 2.2. Note that these conditions are inherent in the mission and should not require special tests for their comparison.

Section 4.0--CRITERIA AND MEASURES

4.1 Criteria.

The following section presents only a sample of the criteria which should be applicable to XM-47 operations. For example, criteria to represent acceptable performance under overload have not been indicated, although they should be, since one of the purposes of the evaluation is to determine that the crew can perform effectively under overload conditions.

The following criteria will be used to verify that system personnel can perform required tasks.

a. Maintenance of an average driving speed in water of 12 knots; in sand, 22 knots; on road surfaces, 35 knots; over fields, 18 knots.

b. Maximum unloading speed: personnel, 3 minutes; supplies, 5 minutes.

c. Probability of hitting a stationary ground target: .95.

d. Probability of hitting a moving ground target: .85.

e. Completion of all procedural tasks within accuracy and time limits specified in applicable TMs.

f. Not more than one discernable instance of incapacitating motion sickness per mission among troops being transported.

g. Ride quality and driving characteristics adequate on all road surfaces.

These last criteria will require quantification before they can be meaningfully measured. It is apparent that much more data must be collected in the personnel performance evaluation than is represented by the preceding criteria, although each datum ideally should be referrable to a specific criterion.

4.2 Measures.

Measures are broken down according to the following categories: performance data, environment, equipment-human compatability, and subjective data. The list of measures is incomplete and illustrative only.

The following data will be recorded:

a. Performance data

(1) Start/stop time for each mission.

(2) Start/stop time for loading/unloading troops/supplies.

(3) Instances of incapacitating motion sickness in troops being transported.

(4) All internal and external communications.

These will require more detailed analysis at a later time; we do not presently have specific criteria for communications analysis.

(5) Start/stop time for premission equipment checkout.

(6) Start/stop time for postmission equipment checkout.

(7) Start/stop time for all instances in which XM-47 attacks a target or is itself attacked.

(8) Number of rounds fired and hits against surface targets.

(9) Number of rounds fired and hits against aircraft targets.

(10) Type and number of instances of error in performing procedural tasks.

Note that without instrumentation it will be difficult for an observer to note all errors; hence only major errors with significant effects on personnel will be recognized and recorded.

(11) Average driving speed on various types of surfaces.

(12) Start/stop time for all contacts with targets.

Where these are very frequent, it may be difficult to get highly precise data.

(13) Start/stop time (downtime) resulting from all equipment malfunctions.

(14) Actions taken to restore equipment to functioning status.

Assuming that observational data recording methods only are available, this information will not be very detailed (e.g., who worked on the equipment, major actions taken).

(15) Distance error in arriving at beach head location.

b. Environment

Noise, lighting, vibration, temperature measures extracted from MIL-STD 1472B (Department of Defense, 1974).

c. Equipment-Human Compatability

(1) Adequacy of equipment layout for operator use, including major human engineering discrepancies.

(2) Aspects of equipment layout and operating procedures which should be modified.

These are more properly recommendations for corrective action. However, since the observations and deductions on which these recommendations are based must be secured during the actual test, they are listed here.

(3) Discrepancy report: All discrepancies from optimal functioning (including equipment failures) will be recorded at the close of the mission.

d. Subjective data

(1) Operator rating of adequacy of environmental conditions and suggestions for modification.

(2) Troop evaluation of ride comfort.

(3) Driver evaluation of driving and ride characteristics of XM-47.

(4) Operator evaluation of aspects which require additional emphasis in training.

(5) Tank personnel evaluation of adequacy with which the total mission was accomplished.

Subjective data collection techniques will be described in Section Three of this manual.

Section 5.0--DATA COLLECTION METHODS

5.1 Data Collectors

Data collectors are emphasized in the Test Plan because, in an operational system test conducted in the field, automatic methods of data collection may not be feasible.

5.1.1 Number. Two (2) data collectors will accompany each XM-47 test mission. The reason for having two data collectors is to check the reliability of the observations made. Should it be determined, on the basis of initial tests, that one data collector will provide reliable data, the number of data collectors will be reduced.

5.1.2 Data Collection Tasks. Data collectors will record the start/stop time of all discrete functions that are described in section 4.2. Where instrumentation is required for recording personnel performance data (e.g., tape recorders for communications), they will activate and monitor the functioning of the instruments. During the test, they will record all significant events by maintaining a running diary (see appended forms). All data recording will be on the basis of noninterference with on-going mission tasks. At the conclusion of the test mission they will administer, collect, and process all paper-and-pencil data forms.

5.1.3 Data Collection Training. It is assumed that the data collectors will have been given factory training on the XM-47. Specific data collection training will take the form of dry runs to familiarize personnel with data collection procedures. It is estimated that at least two (2) dry runs, corresponding as much as possible to the actual test operation, should be made. During the dry run, data will be collected as it would be in the test. Following each dry run, the data collected will be examined and a critique made by the Personnel Performance Data Evaluator.

The nature of the training to be provided data collectors depends ultimately on the functions assigned them. This, in turn, depends at least partially on the system being evaluated. Rarely will a formal training course be required, but a dry run such as indicated above will always be necessary unless the data collection task is extremely simple.

5.2 Data Collection Forms. The following records are appended to this test plan:

a. Time and event records.
b. Instrumentation data records.
c. Subjective data records (questionnaires, rating scales).

5.3 Data Collection Procedures. These have already been described in section 5.1.2 (Data Collection Tasks).

More detailed data collection procedures will be written only when necessary.

5.4 Instrumentation. The following equipment will be installed aboard the XM-47:

a. Sound level recorders.

b. Accelerometers for measurement of motion and vibration.

c. Magnetic tape recorder for continuous recording of crew communications.

d. Equipment for automatic recording of driving speeds.

e. Laser target recorder attached to missile firing equipment for automatic recording of firing and hits.

The Appendix to this test plan presents drawings of equipment locations and attachment details.

Data collection personnel will be responsible for monitoring the operation of the equipment, periodic recording of data values, and performing first-level maintenance should a failure occur.

Failure of instrumentation during the test will not be cause for cancellation of a test mission.

Section 6.0--SUBJECTS

6.1 Number. Subjects are of two types: (a) the XM-47 crew of 3, and (b) troops carried (10).

6.2 Required Characteristics. Selection of the XM-47 crew is the responsibility of the Test Director. Crew members must meet basic qualifications of the following MOS: _____. In addition, they must have satisfactorily completed the contractor's factory training course in operation of the XM-47. They must also be proficiency-certified in firing the Hurricane antitank missile and the Typhoon SAM.

The troops carried in troop-carrying missions need no special qualifications.

Section 7.0--DATA ANALYSIS

The following analyses will be performed:

a. Determination of mission duration.

b. Determination of maximum time for loading and unloading troops and supplies.

c. Determination of pre- and postmission checkout duration.

d. Proportion of ground and air targets detected and correctly identified and compared with criterion.

e. Proportion of ground and air targets hit and compared with criterion.

f. Comparison of average driving speeds per type of surface.

g. Number of equipment malfunctions per equipment and downtime per equipment.

h. Average error in arriving at beach head.

i. Number of instances of incapacitating motion sickness in troops being transported.

j. Determination of mean and maximum temperature during mission.

k. Determination of mean and maximum vibration during mission and comparison with standards.

l. Determination of mean and maximum noise level during mission and comparison with standards.

m. Instances of major error in operation of equipment.

n. Evaluation of adequacy of equipment layout.

o. Evaluation of adequacy of environment conditions and ride comfort.

p. Evaluation of adequacy of driving characteristics.

q. Subject recommendations for modification of equipment design, procedures, and training.

r. Comparison of tank crew performance under day and nighttime conditions.

s. Comparison of tank crew performance in vehicles (1) and (2).

t. Comparison of tank crew performance under troop carrying and supply conditions.

u. Comparison of tank crew performance over succeeding missions (mission (1) vs. (2) vs. (3), etc.).

Note that in these last comparisons the measures to be employed will make use of most if not all the measures listed in section 4.2.

It will be seen by comparison with section 4.2 that the data analysis consists largely of a compilation of the measures in that section. No very sophisticated experimental design is required because this operational test is not concerned with the examination of variables except for the comparison of day vs. nighttime operations, etc. In general, the conditions under which operational system tests are conducted do not permit elaborate experimental designs.

Section 8.0--TESTING SCHEDULE

The personnel performance test schedule is concurrent with the overall test schedule. No special test missions are required for the collection of personnel data. Performance data will be collected during all tests in which personnel operate equipment in an operational manner.

Development of Quantitative Performance Criteria

One of the required tasks in developing the test plan is the specification of personnel performance criteria. Without these, no meaningful evaluation of personnel performance or the system as a whole is possible. Ideally, such criteria should be available in function/task analysis documents and TMs accompanying the prototype system to be evaluated. Unfortunately this is often not the case, and so the resultant evaluation--although providing useful information about the system and how it should be improved--does not accomplish the purpose of the evaluation; i.e., verifying that personnel performance satisfies system objectives. This substantially reduces the value of the operational system test.

Even when quantitative criteria are not available, it is possible to derive them, or at least to develop approximate values of such criteria. The purpose of this section is to describe a systematic procedure for deriving these quantitative values. The procedure described is based in part on what has been termed the Delphi technique.[2] Since performing this procedure will take time and effort, the development of the Personnel Performance Test Plan cannot wait until shortly before the evaluation is to begin.

---

[2]For information on the Delphi technique, see the following: Dalkey, N., and Helmer, F. An experimental application of the Delphi method to the use of experts. Management Science, 1963, 9, 458-467 and Sander, S. I. Delphi: Characteristics and applications (NPRDC Technical Note 76-2). San Diego: Navy Personnel Research and Development Center, October 1975.

At this point, it is desirable to digress and distinguish between overall system criteria and personnel performance criteria--a distinction which was made previously but which bears repetition.

Example: An example of a system output criterion is the following: Accuracy of the artillery piece should be ±50 yards. An example of a personnel criterion is: Accuracy in laying (pointing) the piece should be ±10 mils.

It is important for the evaluator to be able to distinguish between system and personnel performance criteria because he may otherwise fail to measure personnel performance correctly. System criteria describe what is required of the system as a whole, which includes personnel performance; personnel criteria deal only with human functions in the system. In the above example, the measurement of artillery accuracy includes not only the accuracy with which the piece is laid, but physical characteristics of the gun itself, windage conditions, etc. It may appear as if measurements of the system criterion would be sufficient for test purposes, because it includes the effect of personnel performance. However, to fail to measure the human function individually would be to lose much of the value of the system measurement. Suppose, for example, that accuracy of the rifle were actually ±100 yards, which is unacceptable. What is the cause of the inaccuracy?

The major factor causing the unacceptable inaccuracy might be failure of the artillery battery to lay the gun correctly. If one did not measure the human function separately and then relate the human error to the system error, one would never know what needed to be done to bring the error within acceptable limits. One of the major goals of personnel performance testing is to relate personnel error to system performance; one cannot do this except by measuring both factors.

The reason for the lengthy discussion of system vs. human criteria is that system criteria are much more likely to be expressed quantitatively in system documents than are human criteria. As a result the test planner may be tempted to concentrate on the system criteria and to ignore the human ones.

One of the first steps in development of the test plan is to examine the available system data and determine how much criterion information is available. In the case of the system as a whole, the planner should list all the functions to be performed by the system, and then determine for which functions quantitative data are available. He should make a similar list of personnel functions and tasks and the quantitative standards available for these.

Where does this list come from? Theoretically, function/task analysis documents are produced during system development and should accompany the prototype system. Review of these documents should give the test planner at least the major personnel functions and tasks. The documents may not provide a complete listing, however, and the planner will then have to supplement the documents.

The general strategy for deriving a list of tasks together with their quantitative criteria is to make use of a group of subject matter experts. A subject matter expert is someone highly experienced in utilizing a system of the type being evaluated. For example, an officer who has several years' experience in tank operations might well be considered a subject matter expert for a new tank.

There is nothing unusual in the use of subject matter experts to assist in the development of a new system. The Delphi-Like procedure described below is simply a more systematic way of securing certain judgments than might otherwise be secured.

Step 1: The test planner will recruit a number of subject matter experts. Because of the variability inherent in subjective judgments, even from highly experienced personnel, it is best to have as large a number of experts as one can find. Practical considerations enforce an upper limit on the number of these experts. It is recommended that no less than three nor more than five experts be recruited.

Step 2: The test planner will present to the group of experts a list of tasks to be performed by operators of the system to be evaluated. Each expert will be asked to review this list individually. (The reason is that the planner wishes each expert's opinion to receive full weight. In a group discussion, higher-ranking experts may dominate lower-ranking ones; those with more forceful personalities may overawe those with less aggresive personalities.) Clearcut instructions should be written so that each expert knows exactly what he is supposed to do. Each expert will be asked to answer the following questions:

a. Is the list of operator tasks complete? If not, add whatever tasks you consider important to accomplishment of the system mission.

b. Is the level of the tasks described meaningful in terms of accomplishing the task?

Refer back to the previous discussion on this topic. Remember that some tasks may be so grossly described that they include (but do not specify) a number of subtasks for which data should be collected. These latter should be specified. On the other hand, certain tasks may be so molecular that their measurement would be pointless in terms of the overall goal of accomplishing the system mission.

If the task is not described correctly, have the expert rewrite the task as he sees fit, with emphasis on the meaningfulness of measuring that task.

Step 3: After each expert individually has reviewed and modified the list of tasks, collect the revised lists. Then the planner will combine all the lists into one, noting which tasks have been eliminated and which have been added.

Step 4: At this point, reconvene the experts--this time as a group. Give each one the combined task list. Have them examine the list and attempt to agree on that list. If agreement cannot be reached on any task, eliminate that task on the assumption that any such task is probably not of significant importance to mission accomplishment.

Step 5: Once a composite list is agreed upon, ask the experts to take that list and now--again individually--rank-order the tasks in terms of importance. Three criteria of importance ought to be kept in mind by the experts when they rank-order the tasks:

a. Importance in terms of accomplishing the system mission.

b. Importance in terms of system safety (e.g., maintaining the integrity of the system in the face of enemy fire).

c. Importance to the safety of operator personnel.

The reason for prioritizing the tasks in terms of importance is that it may not be possible to measure all of them during testing. The number of data collectors may be too few to collect data on every test or for every task in a given test exercise. Certain tasks may be performed concurrently, which means that two or more data collectors will be required. Some of the tasks may be of such relatively slight importance that there is no pressing need to collect data on them. Of course, if the number of tasks is sufficiently small, and they are all equally important (unlikely but possible), then the step of prioritizing tasks can be eliminated.

At the same time, in the interest of reducing the claim upon the time of the subject matter experts, they should be asked to indicate for each task the quantitiative criterion to which the task should be performed.

> Example: Task (in connection with artillery forward observers (FO): Locate enemy position by reference to map/ grid coordinates. Criterion: ±100 yards.

Selection of a criterion value should be based on operational logic. If, for example, the task is to locate an enemy battery for counter-battery fire, the criterion standard should not be more precise than counter-fire accuracy would demand. If artillery cannot fire more accurately than ±100 yards, there is no point in requiring the FO to locate the enemy target to ±50 yards. On the other hand, if artillery fire accuracy is ±50 yards, one should require the FO to locate to the same value (provided that this is possible within the FO's perceptual capabilities).

Step 6: After the subject matter experts have individually prioritized the tasks and supplied quantitative criteria (qualitiative criteria should not be accepted, since these cannot be used for evaluation purposes), the planner will collect the lists and criteria and develop a composite product.

Step 7: Call the subject matter experts together as a group and have them review the composite list which indicates all the priorities and

criterion values specified previously. As in preceding steps, the experts will be asked to thrash out all disagreements and develop a list on which they all agree. Where the experts cannot agree on a task priority or a criterion standard, take the median value of the judgments made. At this point, the work of the subject matter experts is completed, at least for the job of prioritizing tasks and establishing criterion values. However, these experts may be utilized later in reviewing the data analysis.

Step 8: On the basis of the task/criterion listing and using as a selection criterion the number of data collectors he has available, the test planner will select whatever number of tasks he considers as feasible to measure. The criteria specified for those tasks will be noted and made available to data collectors for use in their measurements.

In complex tasks the task may have multiple dimensions. For example, in air traffic control, the following performance criteria may all be important: (a) distance between aircraft landing; (b) number of aircraft in a landing pattern; and (c) number of aircraft landed in a given time period. The test planner must ensure that subject matter experts do not ignore some relevant task dimensions for which a standard must be available.

In evaluating the importance of the tasks, the planner may wish to have subject matter experts indicate the absolute degree of importance the task has, rather than (or in addition to) its rank order. The expert may be asked to rate each task on a scale such as the following:

Example:

1_____2_____3_____4_____5

1. Critical to system functioning.
2. Extremely important to system functioning.
3. Very important to system functioning.
4. Somewhat important; can be delayed, but should be performed.
5. Relatively unimportant; in an emergency may be skipped without affecting system operations.

The planner may also find certain tasks whose standard of performance depends not on the individual operator but, rather, on circumstances outside the operator's control (e.g., other systems or the enemy's action). Such tasks present special problems in setting up a quantitative performance criterion.

In establishing a criterion for such tasks, the experts will have to estimate the maximum number of inputs to the operator that other systems or the enemy are likely to provide. For example, in establishing the number of coded messages which an intelligence operator would be expected to receive, decode, and transmit to G-2, the expert must consider, on the basis of his experience, the highest likely workload which the operator may have and the maximum amount of work he can be expected to perform under these circumstances. This is not a very satisfactory principle, but it is the best one that can be established for this type of uncertain situation.

The criteria thus established are, of course, not "fixed in concrete"; actual test experience may cause them to be modified.

SECTION THREE--THE SELECTION AND DEVELOPMENT OF MEASURES AND MEASUREMENT METHODS

The measures and methods described in this section can be used for both individual operator and team evaluations. Ideally, the most appropriate measures are objective, quantitative, unobtrusive, and easy to collect, require no specialized data collection techniques or instrumentation, and cost little or nothing. Objective measures are preferable, but often cannot be used. Subjective methods are quantifiable and can be used to supplement objective ones. The objective measures described include reaction time, duration, accuracy, frequency of occurrence, and amount achieved. The subjective methods available include the interview, questionnaire, observation, ratings, checklists, and critical incidents. This section describes the information each measure/method can supply, under what circumstances it can and cannot be used, its advantages and disadvantages compared with other measures/methods, and factors to be considered in selecting the measure or method.

Introduction

The personnel performance test planner has a major responsiblilty to select the measures and measurement methods with which data will be collected. Initially, the test planner selects only a general class of measure (e.g., reaction time); thereafter, he must develop a specific measure by applying the general measure to the particular task and system he is evaluating. For example, he may decide that he wishes to measure an equipment operator's reaction time. He must then decide when he should begin timing performance and the specific response that represents the end of the task period being timed. On the other hand, if he wishes to assess accuracy, he must first define what accuracy (or its converse, error) is in the context of the particular task and system.

Much theoretical material has been written about individual measurement methods and ways of utilizing them. However, since this manual is designed to be used in the actual conduct of an operational system test (OST), only the most important elements of this material have been included; matters of theory which cannot be immediately applied to the OST have been largely ignored. The test planner should, if at all possible, have the assistance of a human factors specialist in developing his measurement methods.

Examples

In order to describe these methods concretely, an actual Marine Corps system--the Tactical Air Communications Central (TACC) (AN/TYQ-1)--will be used as a continuing example throughout the remainder of this section. The following is a short description of this system, its equipment, and the functions performed by its operators. (More detailed information can be found in Marine Corps Technical Manual TM-04428A-1005-14/1.)

The TACC system is part of the Marine Tactical Data System, which is a landbased air control facility. The particular subsystem of the TACC used as the example is the Operations Section, which has five subsections: command, air defense, air support, traffic, and display. It requires 19 operators, although only 11 of these are key personnel.

The major equipments operated by these personnel are five situation display consoles, six communications desks, a Weapons Availability Status Display Group (WASDG), eight manual status boards, and a map board. A computer generates alphanumeric and vector (line) video information for display on the situation display consoles and WASDG, and continuously updates these to provide operators with current tactical information. The situation display consoles contain information needed to evaluate the tactical situation such as a reference map, friendly and hostile aircraft symbols, projected track of aircraft (lines), and reference and identification data (alphanumerics). The operator can display elements at different intensities or remove display elements from the CRT screen. He can select different scales for the display and change the center of the display to focus on a particular area of interest.

The WASDG displays alphanumeric information concerning targets and interceptors (engaged or nonengaged), fire units available, and miscellaneous friendly resources. This is for group viewing.

The Operations section also includes a total of 10 intercommunication stations with two-way voice communications between operations and planning group personnel. These communications desks interface with remote field telephones and radio. A teletype permits secure or nonsecure teletype communications with remote facilities; it provides relative low-speed, textual traffic. A loudspeaker-amplifier permits public address of any voice communications signal available at the communications desk.

TACC personnel perform a wide variety of operational and behavioral functions. The operational functions include the following:

1. Monitor, supervise, and coordinate the control of aircraft for air and assault support.

2. Control the launch and allocation of on-call air support aircraft.

3. Coordinate all air traffic in area of responsibility.

4. Monitor equipment status and operational posture of other Marine Corps organizations relative to air support.

The behavioral functions include:

1. Detection of video signals.

2. Operation of control consoles.

3. Analyses of alphanumeric and symbolic data.

4. Verbal communication.

5. Interaction with others.

Although operations are governed by superordinate procedures, they are somewhat unstructured at the working level, flowing from immediate needs.

The configuration of Operations section stations in their inflatable shelter is shown in Figure 3-1.

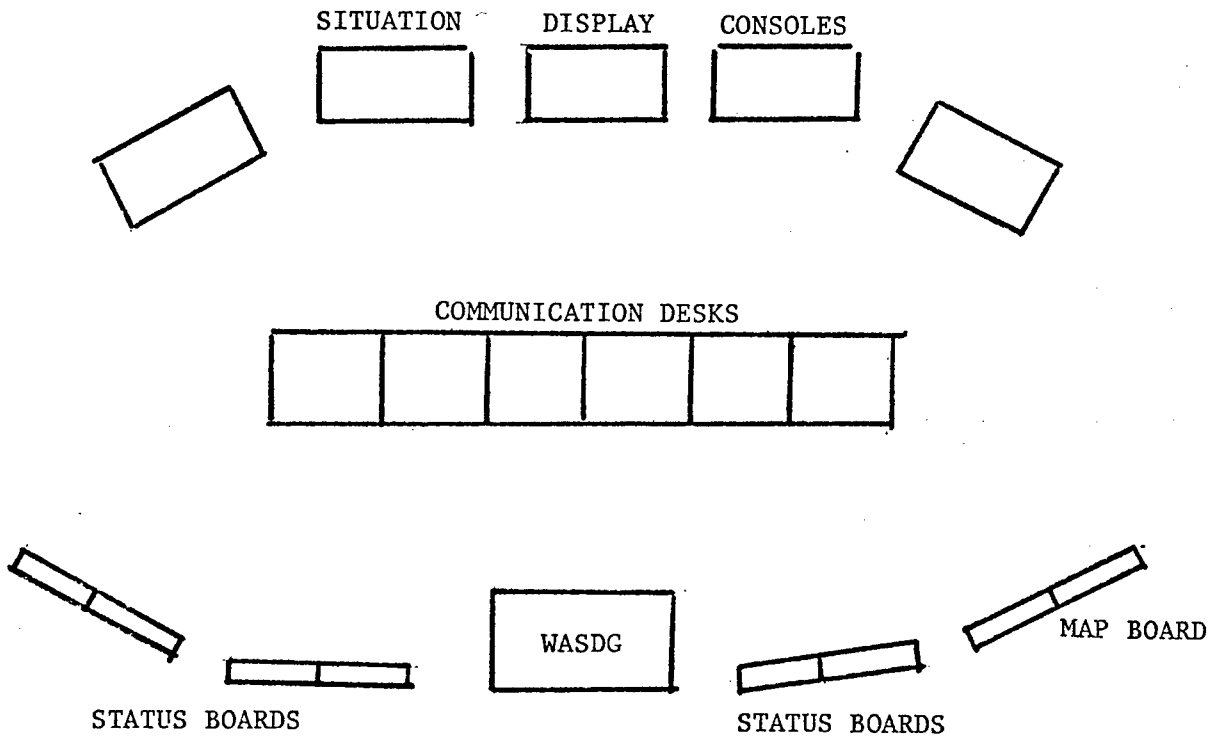


Figure 3-1. Equipment arrangement of Operations section personnel.

Criteria for Selection of Measures and Methods

The evaluator should apply the following criteria in selecting and developing his measures and measurement methods:

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 some inaccuracy and inconsistency. Human data collection requires (a) recognition of the behavior to be reported (i.e., its occurrence), (b) determination of the relevant characteristics of that behavior, and (c) recording of the relevant data. Since considerable interpretation is involved in each of these phases, they are subject to error.

The previous paragraph began with the word "ideally." As a matter of practicality and cost, many measures employed in OST cannot be completely objective and, as will be seen subsequently, it would be undesirable to restrict measures to those that are completely objective.

2. Quantitative. Quantitative measures can be scaled and combined with other quantitative data; this is not true of qualitative data. Although qualitative measures can be useful to support or to explain the meaning of quantitative data, they cannot stand alone.

3. Unobtrusive. The act of gathering data should not affect the manner in which the operator performs his tasks. All data collection agencies (physical as well as human) should ideally be invisible to the performer. If operators become unduly aware of these agencies, they may perform in ways that are not representative of their operational activity. Operators may try too hard, they may make errors because of anxiety, and they may deliberately distort their behavior to fit some preconceived notion of what the evaluator wants. Since the goal of the OST is to verify operational performance, anything that leads to nonrepresentative behavior reduces test validity.

Test managers and operational commanders also generally object to any interference with ongoing tasks. An obtrusive data collection method (one which is highly visible) may be viewed as interfering, even if it does not hinder operational activity.

4. Easy to Collect. Any measure whose implementation makes extensive demands on the capability of data collectors is likely to produce errors in the data gathering process. If the data collector is required to make a difficult perceptual discrimination or computation, he may well make an error, thus reducing the validity of the resultant data. If, for example, he were required to note the precise time at which each signal on an air surveillance radar appeared, he would quickly lose track.

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 and, although some training should obviously be given them, it should (for time and cost reasons) be as little as necessary for them to perform their jobs. More important, special data collection techniques are likely to make it impossible to employ Marine Corps personnel as data collectors, because they will probably lack the needed technical background. It is always desirable to utilize operational personnel as data collectors (if data collectors are required) because (a) their familiarity with the task being evaluated may improve the precision of the data they collect, (b) operational personnel are less likely to be viewed by task performers as obtrusive elements, and (c) data collectors with the necessary technical background are likely to be scarce, and their unavailability may constrain data collection.

6. Require No Specialized Instrumentation. If the OST is being conducted in a field environment, specialized instrumentation may not function too well in that environment. Such instrumentation is often too delicate for the rough usage it may get in the field and therefore tends to malfunction more often than simpler equipment. 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 for not conducting 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 are of course ideals, and in the real OST world it is often impossible to satisfy these criteria completely. It may be necessary to use qualitative, subjective, or obtrusive measures or special data-gathering techniques, much as one would wish otherwise. The reason for listing these criteria is to provide a standard at which the evaluator can aim, rather than to make them hard and fast rules which often cannot be followed.

Available Measures and Methods

A listing of available objective and subjective performance measures and methods for use in evaluating both individuals or teams is provided on pages 3-7--3-9.[1] These measures and methods are discussed in the following pages organized according to the following outline:

1. Definition.
2. Information provided (by the method).
3. Use factors (factors to be considered in using this method).
4. Problems (in using this method).
5. Example.
6. Summary (of steps to be followed in using this method).

[1]Modified from a listing published by Smode, A. F., et al. The measurement of advanced flight vehicle crew proficiency in synthetic ground environments (MRL-TDR-62-2). Wright-Patterson Air Force Base: Behavioral Sciences Laboratory, February 1962.

LISTING OF AVAILABLE PERSONNEL PERFORMANCE MEASURES AND METHODS

OBJECTIVE

1. Time

a. Reaction time (i.e., time for an operator to perceive an event or start an action in response to some initiating stimulus).

b. Duration (i.e., the total time required for a task to be completed). In tracking a target, it is percent time on target.

c. Time between events (e.g., mean time between failures).

2. Accuracy

Accuracy in:

a. Observing and identifying stimuli or occurrences (internal or external to the system).

b. Estimating distance, direction, speed, time of movement of objects.

c. Detecting a change in events or stimuli over time.

d. Recognizing a signal in a noise or high target density background.

e. Recognizing an out-of-tolerance condition.

f. Positioning a control or a weapon or using a tool.

g. Reading displays.

h. Selecting among alternatives (e.g., decision making).

i. Making a series of discrete responses (e.g., throwing switches).

j. Making a series of continuing responses (e.g., tracking).

k. Communicating.

l. Error analysis, in terms of:

(1) Amplitude

(2) Frequency

(3) Type

(4) Changes over time

3. Frequency of occurrence

a. Number of responses made by operator in performing tasks such as the following:

(1) Observing.

(2) Controlling and manipulating.

(3) Communicating (e.g., verbal or written reports, requests for information, etc.).

(4) Personnel interactions.

(5) Maintenance diagnostic checks.

b. Number of effects of task performance, such as the following:

(1) Number of errors.

(2) Number of out-of-tolerance conditions.

4. Amount Achieved

a. Cumulative responses (i.e., total number of responses made).

b. Degree of success in accomplishing tasks and mission.

c. Achieved reliability (percent of tasks accomplished, or ratio of tasks successfully accomplished to all tasks undertaken).

d. Proficiency scores (e.g., written test performance).

e. Terminal or steady state value (e.g., a temperature high point).

f. Performance variability (e.g., degrees change per hour).

5. Amount Consumed

a. Physical resources consumed in terms of activity or time, such as:

(1) Fuel/energy consumed or conserved.

(2) Units consumed in accomplishing tasks (e.g., weapons or ammunition expended).

b. Personnel resources consumed in terms of casualties.

c. Man-hours consumed (response time x number of personnel involved).

SUBJECTIVE

Available subjective measures are:

1. Interviews.

2. Questionnaires.

3. Observations.

4. Ratings.

5. Checklists.

6. Notation of critical incidents.

These measures are used for the following purposes:

1. Performance Efficiency Determination.

a. Performance of the operator/team (ratings).

b. Success of the task/mission (ratings).

c. Adequacy of equipment, procedures, logistics, technical data, and training (checklist, interview, ratings).

2. Factors Affecting Performance.

a. Identification of factors affecting performance (interview, questionnaire, ratings).

b. Attributes possessed by task and performance (e.g., degree of comfort) (ratings).

3. Event Occurrence.

a. Description of task performance (observation, interview).

b. Unusual occurrences (critical incidents).

In the previous section, we said that an operational system evaluation has the following goals (as these relate to personnel performance):

1. Verification that the system (including its personnel) satisfies system requirements.

2. Prediction of the future operational performance of the prototype system under evaluation.

3. Diagnosis of problems found as a result of testing the system.

4. Provision of data that may be applied to the analysis of current and future systems.

5. Identification of the personnel resources needed to utilize the system effectively.

6. Provision of data for future training of personnel who will operate the system being evaluated.

Table 3-1 categorizes each measure/method listed above in terms of its potential use in the evaluation for meeting these goals.

Objective Measures and Methods

Time

Reaction Time.

1. Definition. Reaction time (RT) is the time between the occurrence of an event requiring an action on the part of the operator or team (the initiating stimulus) and the start of the action (the RT response) demanded by the event (see Figure 3-2). Since the initiating stimulus must be recognized by an operator, that stimulus is likely to be something observed directly by him or displayed on a CRT or indicator display. The RT response is the operator's action in throwing a switch, making a verbal report, etc.

2. Information Provided. Table 3-1 indicates that the RT measures can be used to verify (or fail to verify) that personnel can meet system requirements and that the data can help to predict the operational performance of the new system as well as to provide data useful in planning future systems of the same type.

The major purpose of measuring RT is to determine how quickly the operator/team can react to an initiating stimulus. Before selecting this method, the test planner must ask whether this information is necessary for the evaluation.

Where the operator or team is required to respond to an initiating stimulus in a fixed (minimal) period of time, his or its ability to do so must be verified. If the RT required of an operator is very short, he may have difficulty accomplishing the task. If one examines task requirements in advance of evaluation, and required RT appears to be quite short, then RT measurement may determine if the operator's capabilities are being exceeded.

If the operator/team is required to react as quickly as possible (even though a fixed RT is not specified), the test planner will wish to determine the minimum and maximum RT he can expect of the system. However, this information cannot be used to evaluate the effectiveness of the system unless an RT criterion (standard) exists or can be developed. If an RT criterion is desirable but does not exist, the operational personnel who must operate the system can develop such a criterion by the method described in Section Two. If a time requirement does not exist, either explicitly or by implication, RT is unlikely to be of value in evaluating the system.

The nature of the system will determine whether or not an RT measure is meaningful. RT is significant only when the speed of a reaction to the initiating stimulus will determine the effectiveness of the system response.

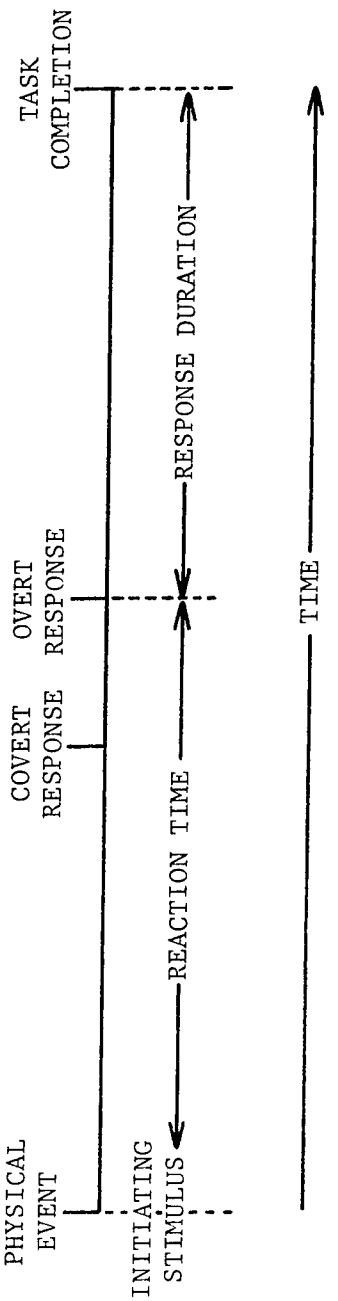


Figure 3-2. Reaction time and response duration measures.

Table 3-1

Information Provided by Measures and Methods

Use	Objective					Subjective					
	Time[a] Measures	Accuracy	Frequency	Amount Achieved	Amount Consumed	Inter-views	Question-naires	Observa-tions	Ratings	Check-lists	Critical Incidents
Verification of System Performance	X	X	X	X	X			X	X	X	
Prediction of Operational Performance	X	X	X	X	X			X	X		
Diagnosis of Problems	X	X				X	X	X		X	X
Data for Future Use	X	X	X	X	X			X	X		
Identify Personnel Resources Required						X	X	X			X
Data for Future Training	/	X				X	X	X	X		X

[a]Reaction time, duration, and time between events.

Beyond the verification that personnel requirements have been satisfied, RT can be useful in suggesting the speed with which the new system, when deployed, will respond. This obviously impacts upon its ability to accomplish its operational goals. Moreover, if a new advanced system of the same type is to be procured in the future, the RT data will supply information with which the performance of the new system can be compared.

3. Use Factors. A distinction must be made between operator and team RT. The team RT may be distinctly different from that of the individual. An officer in charge of an artillery unit may respond to the sight of an enemy target (the initiating stimulus) by issuing a command to his battery to fire (the officer's RT response). The officer's command then becomes the initiating stimulus for the battery (team) and the firing of the first round becomes the team's RT response.

In many cases, the determination of operator RT may be much less important than team RT. The individual operator's RT is, in many cases, almost immediate and thus may have minimal effect on system RT (except where the operator is in effect the system, as in a single-pilot aircraft). Team RT may be much longer, however, because of team member interactions that delay the team response and may have significant impact on system RT.

Once the team has been presented with the initiating stimulus, the RT of any individual in that team need not be measured. To the extent that team RT depends on the individual RTs of its members, the latter is included in the former. Whether one measures individual or team RT depends on which is being evaluated. If the team RT is excessive (i.e., greater than that allowed by the criterion or whatever operational commanders feel is excessive), it may be necessary to measure the RTs of individual team members to determine the source of the difficulty.

If RT is important, it must be measured precisely. Where short time intervals are involved (fractions of seconds, although in most cases no shorter than seconds or minutes), RT measurement may require the use of highly sophisticated instrumentation, particularly in the case of operator RT. Team RT is usually somewhat longer and may not require such precision; a stop watch may be sufficient.

In collecting RT data, a preliminary analysis should be made of all the operations involved in the task to be measured. This means plotting out each step in the task in terms of the time it should take to perform the step, including required RTs to initiating stimuli.

4. Problems. The recognition of the initiating stimulus may be difficult. Where the stimulus cannot be easily anticipated by an observer, when it is accompanied by irrelevant stimuli from which the observer must distinguish it, or when the initiating stimulus is itself very short, it may be difficult for both the operator and the RT data collector to recognize that stimulus. Suppose, for example, that one wished to determine the operator's RT to the appearance of an enemy aircraft on a radar scope. It is possible for the data collector to note when the echo first appeared on the scope. However, the observer's RT to the appearance of this stimulus is much the

same as that of the operator whose performance he is measuring; that is, it is compounded of the time needed to detect the blip, to distinguish it from clutter, to note whether the operator has detected it, to throw the switch on the timer, etc. This means that the observer's error in recording this form of RT will be equal to the RT for the operator. This type of situation demands such precise time recording that often only special instrumentation has the needed capability.

Another problem is the nature of the RT response. In situations in which the operator's response is covert (i.e., perceptual or analytic and not expressed as a control manipulation), it is extremely difficult for an observer (or for that matter, for instrumentation) to pinpoint exactly when the response has occurred. In surveillance systems, the operator may initially respond to the appearance of a target by continuing to observe it. Unless he is required to report verbally or to throw a switch indicating that he has recognized the target, there will be no observable indication that he has in fact recognized the target. In situations involving perception, monitoring, silent analysis, etc., there is no overt response. In such cases, it becomes necessary to wait until the operator activates a control and to consider this activation as the RT response.

5. Example. A major function of the TACC is to control air space over a specified area of responsibility. This requires the Air Defense Coordinator (ADC), for example, to detect and identify on his CRT any aircraft entering his air space. Since the aircraft may be unfriendly, the more rapid the detection and identification, the more efficient the system.

The appearance of a blip on the CRT is the initiating stimulus for the ADC. His RT response is his perception (recognition) of the unidentified aircraft. Since this recognition is covert (i.e., not observable or measurable without extremely elaborate instrumentation), a more feasible RT response would be his "hooking" the aircraft (a control action) to secure additional information about it.

The most important question in the selection of an RT measure for TACC evaluation is whether such a measure is meaningful in the TACC context. Performance standards for TACC operations do not include a quantitative RT requirement although, obviously, RT should be as short as possible. Since there are five operators controlling various aspects of the air space, it would seem reasonable that, if RT were collected for one operator, it should be collected for all. However, one might reduce the magnitude of this task by selecting only critical personnel ("key men") and measuring only their critical functions.

Moreover, the operator performs several sequential functions for which an RT measure should be secured if RT is to be measured at all. For example, the ADC may have to make an unscheduled launch of an aircraft to intercept an unfriendly aircraft and, obviously, the time between identifying an unfriendly and launching an interceptor is critical.

With a system the size of TACC, one can select many RT measures. If the TACC had the capability of automatically recording the time at which

each control was activated and/or of recording verbal communication on a time base, the problem of measuring RT would be much reduced, although the data analysis of the automatic printouts would still be considerable. (Such an automatic monitoring system is in use aboard ship with the Navy Tactical Data System.)

The TACC represents an evaluation situation in which RT measurements would be highly desirable but in which the complexity of the situation, without automatic instrumentation, makes this measurement difficult. The absence of a performance criterion makes interpretation of the resultant data also problematical, because what would any particular RT value mean? If it takes the average operator 6.8 seconds from the first appearance of a target on the screen to "hooking" it, is this good or bad? Again we see that the critical factor in this measurement, as in other measurements, is the existence of a criterion.

However, if it is unpromising to attempt to measure the RT of individual operators, is it possible to determine a total TACC RT? No, because each of the operators performs different functions concurrently, so that system activity is a composite of many actions which cannot be meaningfully combined, as they must be if one attempts to secure system RT.

6. Summary. The following steps should be followed in selecting an RT measure:

a. Analyze system requirements to determine whether an RT measure is needed, will be useful, and is feasible in evaluating personnel performance.

b. If the answer to item (a) is yes, analyze task operations to determine:

(1) Whether individual or team RT is needed.
(2) What the initiating stimulus is.
(3) What the RT response is.
(4) The measurement precision required.

c. Determine whether RT can be measured by observation (data collector) or requires instrumentation.

d. If RT is to be measured by observation, ensure that the data collector is aware of what the initiating stimulus and the RT response are and require him to practice the RT measurement.

Duration.

1. Definition. Duration is the time spent accomplishing a task, from the time task performance is triggered (the initiating stimulus in RT measurement) to the time the task is completed (Figure 3-2). This type of measure is an extemely common one. It is often recorded even when there is no apparent need to verify a duration requirement.

2. Information Provided. Duration is important when the system prescribes a maximum duration for a task or group of tasks. If a task must

be performed in no more than 5 minutes, for example, it is important to measure task duration to determine if that requirement is satisfied. The question is one of operator/team capabilities: Can the task(s) be performed in no more than 5 minutes, or does this requirement impose a demand on the operator/team that is physically not possible or, as a consequence of which, accuracy is degraded? An example of critical task duration is often found in maintenance operations where a maximum time for restoring a malfunctioning equipment has been assigned.

Even if a maximum duration is not specified, it is often of interest to the test planner to determine performance duration: a performance duration that is considered excessive by operational commanders will degrade system effectiveness. If, during the OST, difficulties arise in completing a function, duration measurements could be useful in suggesting the cause of the difficulty: a particular subfunction which takes excessively long may be the factor causing the difficulty. Duration data also are useful in predicting the operational performance of the new system and in providing a basis for comparing the performance of future systems with the present one.

If system operations are flexible (as in the TACC) and the operator is performing several functions concurrently or sequentially, the test planner may wish to determine the percentage of his time the operator spends on individual functions (or the time spent performing the task and the time occupied by equipment operations). If an operator, for example, has to receive telephoned messages, to record them, to enter the data into a computer system, etc., the percentage of time he spends on each of these functions may indicate if he is being overloaded by a particular activity.

3. Use Factors. Ordinarily this type of time measurement does not have to be extremely precise, unless there is some system requirement which necessitates highly precise measurement (e.g., as in timing a foot race). When duration is relatively gross, it is easily recorded. For very precise measurements, instrumentation may be required.

Because most duration measurements usually need not be overly precise, the measurement operation is comparatively simple, often demanding no more than a stop watch and a single observer. Where several concurrent tasks must be individually measured, however, several data collectors may be required.

Measurement of team task duration should pose no more problem than measurement of individual task duration. A team task is one in which the individual functions interact to secure a common output. If, as in the case of TACC, distinctly different functions are performed by individual operators, one cannot speak of the total number of these functions as a team task.

The scope of the team effort whose duration is to be measured will obviously be greater than that of any individual task, and it will be necessary prior to measurement to define that scope. As in the case of RT, the duration of the individual tasks comprising the team effort need not be measured unless duration either is or is expected to be excessive.

4. Problems. One possible problem is the precise definition of the unit of performance whose duration is being measured. If duration is measured from the start of system operation to its conclusion (when the system is secured or goes "off the air"), it is relatively easy to record. The measurement of individual task duration (i.e., one task nested within a series of tasks or several tasks performed concurrently) is more difficult. Tasks in a series flow into each other and there may be no clearly defined start and stop point to bound the limits of the measurement. This is another reason for analyzing task characteristics prior to measurement.

The duration of system operations may not be the same as the duration of tasks performed during those operations. Tasks may be performed prior to the equipment taking over, or the operator may have to wait while the hardware is performing its part of the operation. Measurement of the system operation as a whole will therefore include the task, but will not necessarily spell out how long the task took.

5. Example. Once TACC is set in operation, it functions on a continuing 24-hour basis. System duration for TACC as a whole would therefore not be very meaningful. On an individual operator and function basis it would, however, be quite feasible to determine duration. One might, for example, determine the length of time it takes to intercept an unfriendly aircraft; the duration would start from the launch of an interceptor to the time the interceptor is within firing range. From an evaluation standpoint, this datum might be of little value, but it might be significant for diagnosing the decision-making efficiency of the operator, since intercept duration is partially determined by the vector he gives the interceptor pilot.

Since TACC personnel perform many functions, it may be of some interest to determine the distribution of time among these functions (e.g., the percentage of time spent coordinating resources as against the time spent launching aircraft). This information, too, would be primarily of diagnostic value.

6. Summary. The following steps should be followed in measuring duration:

a. Determine whether a duration measure is needed.

b. Determine the unit of performance whose duration will be measured.

c. Determine whether duration should be measured by instrumentation or manually.

d. If duration is to be measured by observation, ensure that the data collector is aware of the start and stop points of his measurement.

Accuracy.

1. Definition. Accuracy, or its converse, error, is probably the most common and perhaps the most useful measure of personnel performance.

There are systems and tasks in which reaction time and duration are not important but accuracy is critical in all. All personnel performance assessment depends wholly or in part on accuracy measures.

Except in those rarely found systems that permit no error (and it is hard to think of an illustration), one would expect some errors to occur. The problem then becomes one of evaluating the significance of the number and type of errors to system performance. Each error is not equivalent to every other error. Even if it were, the relationship of a particular number of errors in performance to the system's capability to accomplish its functions depends on a criterion of the maximum number of permissible errors. For example, suppose that in a new aircraft crewmen made an average of .75 errors in performing a preflight checkout of 38 steps. Does this mean that the checkout procedure for this aircraft was ineffective or that personnel could not perform preflight checkout as required? One cannot answer this question unless the evaluation begins with a standard that N errors of a given type are or are not acceptable. The figure of .75 errors is, therefore, just an interesting statistic.

Any inadequacy in performance can be considered an error, but errors generally consist of incorrect commission or ommission. Errors of commission include performing a nonrequired action and performing a required action incorrectly or out of its required sequence. Errors of omission consist of failing to perform a required action. Errors may occur when personnel do one or more of the following:

a. Utilize data from their equipment (e.g., reading a temperature gauge incorrectly or failing to read it when required).

b. Fail to take action demanded by the equipment (e.g., failing to take emergency action when meters indicate the need to do so or failing to report a potential malfunction).

c. Utilize the environment or objects in the environment (e.g., failing to recognize a geographical landmark or an enemy target).

d. Fail to take action demanded by data describing other systems (e.g., failing to comprehend intelligence on enemy movements).

e. Estimate distance, direction, speed, time (of enemy movements) or make projections of action required on one's own equipment or personnel in the future.

2. Information Provided. Accuracy data are critical to verify that personnel can perform required tasks. Obviously, an inability to perform adequately is manifested by error. To the extent that the system performs during OST as it will perform operationally, accuracy data serve to predict operational performance. Accuracy data assist in the diagnosis of problem areas that arise during OST. The nature of the error, who makes it and when, may reveal an inadequacy in system design that impacts on performance. Accuracy data suggest where training emphasis should be placed, since errors reveal functions that must be strengthened through additional or different

training. Accuracy data are also applicable to future systems since, like RT and duration, these serve as a baseline against which to compare the performance of more advanced systems.

3. Use Factors. It is necessary to know not only error frequency, but also the type of error made. The nature of the error is potentially diagnostic of a situation that needs modification; it may cast light on what needs to be done.

Should an error that is corrected by the operator be counted in the same way as one that is uncorrected? No. Many more errors are corrected than are not; if this were not so, it is unlikely that any equipment could be operated effectively. An error corrected by the operator would not be counted unless it degrades task performance significantly until corrected.

The evaluator also needs to know the criticality of the error. Some errors have potentially significant or catastrophic effects on task and function accomplishment; others do not. For example, an error in performing a continuous function such as tracking may be much more significant for performance than one in a discrete task like throwing a switch, where the error is more visible and can more readily be reversed.

In general, errors that have only a minor effect on performance and that are readily corrected need not be counted in the determination of personnel effectiveness. The more significant the potential effect of an error, the greater weight it should assume in the evaluation. It is possible to weigh errors on a criticality scale such as the one below.

Rating	The error has	As a result
4	Extremely grave effects.	Task cannot be accomplished.
3	Great effect on performance.	Task is accomplished, but inadequately.
2	Moderate effect on performance.	Task is accomplished adequately but is markedly delayed.
1	Slight effect on performance.	Task is accomplished and with only slight delay.
0	No effect on performance.	Task is accomplished adequately and with no delay.

Assume one error rated 3, one rated 1, and three rated 0. The total error score assigned to a particular performance would be 4. Such a technique permits one to compare performances on different tasks. However, it does not provide an absolute evaluation (e.g., the probability of this task being performed correctly in future system operations is .75).

In evaluating the potential consequences of an error, the evaluator must also differentiate between the effect of the error on the individual task and its effect on overall system performance (where more than one task is involved). Some errors impede the performance of the individual task but have little effect on overall system performance, either because the task is not critical to that system performance or because there are compensatory mechanisms in the system which cancel out the error effect. Obviously, a task which, if performed incorrectly, would seriously jeopardize system mission performance would be much more critical than one which affected only the individual job.

The criterion of performance accuracy (which should be defined prior to measurements) should therefore indicate not only the number of allowable errors, but also the type of error which is to be considered in making the assessment and its criticality. Knowing the types of errors that can be made will make it easier for the evaluator to recognize when that error occurs in observing performance. Without such a prior analysis, it may be difficult for the data collector to recognize performance inadequacies. Of course, prior experience with a similar task or system may compensate partially for lack of such an analysis.

Errors can be recorded manually or by instrumentation. Instrumentation is necessary when the erroneous responses: (a) are numerous or frequent, (b) change frequently, (c) have a very short duration, or (d) are difficult to observe directly (e.g., physiological reactions). Manual error recording (by observation or self-report by the operator whose performance is being evaluated) is necessary when (a) considerable interpretation of the response is required in order to determine that an error has in fact occurred, or (b) the error response is recognizable only by the context in which it occurs (machines have difficulty recognizing context). Errors may be recorded manually when (a) operator responses are not overly many or frequent, (b) have a reasonable duration, or (c) are readily recognizable. Statements such as "Not many or frequently" are rather general; their specific values depend on the individual data collection situation. However, even slight experience with the operator responses to be recorded will permit the test planner to make a decision on these points. The tradeoff between automatic and manual data recording of accuracy data is often determined by cost. Test managers prefer not to spend money on instrumentation if they do not have to.

The test planner almost always has a choice about the data he can record. Most systems involve various types and levels of personnel responses and the test planner may elect to record only some of these. Because a choice is possible, it is essential that the test planner specify in advance of the evaluation precisely the kind of data he will collect. Whether data are collected manually or by means of instrumentation, the operator's response must be precisely defined. If it is not, it may be difficult for the data collector to recognize the occurrence of the error.

The previous section recommended that the test planner record measures on those task levels appropriate to the evaluation questions to be answered (e.g., recording the number of target hits rather than a rifleman's finger-squeeze pressure). Within that general constraint, as much error data as

possible should be recorded because varying conclusions are possible depending on the particular errors made.

Should all errors be recorded? The answer is no, and this ties into the preceding discussion of error criticality. Errors that can have no impact on system performance can be ignored (assuming that they are recognized during data collection as having no impact); all other errors should, of course, be recorded. Theoretically, if all personnel actions are relevant to system performance, all errors should be significant. However, in actual practice, it often turns out that certain of these are trivial. The preevaluation analysis of the data to be collected should include, along with a review of the operator's task and the kinds of performances that may occur, an analysis of the effects of these performances.

The determination of what constitutes an error presents special problems when the operator's performance is covert. (See the previous discussion on reaction time.) By this we mean that the task is being performed inside the operator's head. Such tasks are likely to be perceptual (e.g., monitoring a display of some sort) or cognitive (e.g., making a decision or a mental calculation). In such cases, the test planner should look for some overt response which is associated with the covert one. For example, if the operator must throw a switch or make a verbal report based on his covert activities, it is throwing the switch or reporting verbally that become the types of data whose accuracy is recorded. The preevaluation analysis of data to be collected should indicate both covert and overt responses.

Manual techniques for recording errors include the following:

a. If all (or most) potential errors can be categorized in advance of observation, one could develop a checklist of these errors and simply check them off as they occur during task performance. However, this procedure is not recommended for two reasons: First, it is difficult to anticipate all errors. Second, simply noting that an error of a given type has occurred does not supply all the information desired. One wishes to know also at what step in task performance the error has been made and what the operator does (or does not do) as a consequence of his making the error.

b. One could attempt to record all of the task actions exhibited by the operator and later sort out those actions that are erroneous. However, such complete data recording is difficult for an unaided observer; he would need something like a videotape camera. The advantage of comprehensive performance recording (such as one could have from a filmed record) is that it provides a complete record for later analysis, permitting the evaluator to examine the performance more leisurely. This procedure is not recommended, however, unless instrumentation is available.

c. The most common procedure is to use an operating procedure as a sort of "template." As the operator performs his task, the data collector checks off each action on the procedure. A performance deviating from that specified in the procedure would be noted at the appropriate step in the procedure (Figure 3-3).

Task Step	Error and Consequence
1. Turn power switch to ON position.	
2. Read external temperature.	
3. Align power output reading as required by external temperature.	
4. Check CRT gain intensity by depressing GAIN switch and read value displayed on GAIN display.	
5. Record gain value.	

Figure 3-3. Use of operating procedure as error data collection device.

Difficulties arise, however, where the task to be performed is so unstructured (i.e., involves so many contingencies) that a step-by-step procedure for that task cannot be developed. Although one could still use an operating procedure, it would be more general than the one shown in Figure 3-3. Instead of listing discrete steps, the test planner could list the general functions to be performed (e.g., launch aircraft, coordinate assault) and, below these, the major actions required to perform each function. Thus, in vectoring an interceptor to the vicinity of an unfriendly aircraft, the operator has to communicate with the interceptor and make a number of course corrections. Any errors associated with each of these actions would be recorded. Naturally, the more unstructured the task, the more the observer/data collector must know about the job being performed so that he can recognize an error occurrence.

As was pointed out previously, analysis of the errors made, their frequency and amplitude, and where in system operations they occur suggests the potential causes of the error. The type of error is important, since it often reflects the cause.

Basically, however, the evaluator wishes to determine the efficiency of operator performance. This can be done by contrasting the number of tasks successfully accomplished (s) as against the number attempted (n). The ratio (s/n) provides a percentage value which can be interpreted to indicate the future likelihood of these tasks being successfully accomplished. For example, if 10 tasks are attempted (n) and 9 are correctly performed (s), the efficiency ratio if .90. The emphasis here is on successful accomplishment of a task; if errors are made but are retrieved (or even if not retrieved) and the task is completed, it must be recorded (despite errors) as successful. This relates to the criterion of what constitutes an error. One is not simply counting errors; one is trying to determine whether the task was successfully performed.

The determination of successful task accomplishment is essentially a judgment made by the observer/data collector. That judgment is based on the output of the task performance when it is completed. For example, an air intercept is judged as successful when the interceptor is vectored to a point in space where the pilot can report viewing the unfriendly aircraft.

4. Problems. Many of the problems encountered in applying accuracy measures have been discussed as part of the preceding section.

Special problems may arise, however, in measuring team accuracy as distinct from individual operator accuracy. This can be done without excessive difficulty provided there is a product of the team activity. Assume a four-man team in which a decision (e.g., to reconnoitre a particular area) must be based on inputs from three of the men (the fourth man being the decision maker). The accuracy of that decision can obviously be measured. The decision is a team output (not merely the result of the decision maker's activity) because the decision could not be made without the three inputs.

One would not wish, however, merely to examine the team output (decision) without considering the individual performances that led to it. If the decisions were invariably correct, then perhaps one could ignore the individual performance, assuming that they were correct because the output was correct. However, decisions or other team outputs are sometimes incorrect; when this occurs, it is necessary to examine the individual inputs that led to the team output. From the standpoint of the accuracy or quality of the team output, it is essential to measure individual contributions as well as the team output.

A distinction must also be made between an erroneous output and an error made in arriving at that output. As has been pointed out previously, many errors are inconsequential. However, an erroneous output is always significant since it directly impacts on the mission of the team. Hence, a team output measure (when available) should always be gathered.

5. Example. One of the factors that will affect the collection of accuracy data in the TACC is that there is a variable workload (e.g., the need to detect and identify aircraft and to launch interceptors, depending upon enemy incursions). One would therefore expect TACC personnel performance to vary with that load. Thus, TACC performance should be sampled at various times to record performance as a function of the range of loads from very light to very heavy. If the evaluation is performed using "canned" stimuli (e.g., targets produced by a target generator), then that load can be specified in advance. If the evaluation is not performed according to a prepared scenario of inputs to the system, then data collection will depend on targets of opportunity. In either case, since workload for the TACC does ordinarily vary, data sampling is necessary.

The number of activities performed in TACC lends itself to a variety of error measures. There is no question that accuracy measures should be taken, but which ones and how? There is the rather specific procedure employed to operate the situation display consoles--such things as depressing

the correct buttons to "hook" the target or to select a piece of video, etc. The number of errors made in operating the situation display console could be determined without difficulty by observation of the operator's activities. The data collector would have to know its operation since the procedure is so flexible (depending upon the nature of the inputs) that it is impossible to follow a written procedure exactly.

Other operator activities are more difficult to record. Analysis of the functions of TACC personnel reveals that there is considerable decision making, supervision, and coordination. Part of the time they are monitoring the buildup of new tactical situations; much of this activity is covert or, if not covert, at least it cannot be tracked with the aid of a step-by-step procedure. Under these circumstances, it will probably be necessary to make use of ratings of performance quality, either by supervisors or by a knowledgable observer. A more detailed discussion of the form such ratings could take will be given in the section on ratings.

The great variety of activities performed by 11 or more personnel requires some degree of selectivity on the data collector. Even if it were feasible to have one observer for each TACC team member (and that is unlikely) when input loads are high, data collectors will not be able to keep up with the flow of activity. It therefore is necessary to preselect those "key" activities which are most representative of and critical to TACC operations and to concentrate on these. Hence, prior analysis of the functions and tasks performed will be necessary. This analysis should indicate also the kinds of errors that are most likely to occur; this information will help to cue the observer to what he should look for.

Another consequence of the size of the TACC team and the variety of their activities is that instrumentation is not a very feasible way of collecting error data. The covert nature of TACC activities such as coordination forbids their instrumented measurement. To replace an observer, one would need at least three or four video cameras and even these would not provide the degree of detail needed, because it would be difficult to get the cameras close enough to record adequately without their becoming intrusive. And then, of course, the posttest analysis of this material would be a horrendous task. This is a case where the flexibility of an observer, assuming he is properly trained for the data collection task, makes him much more valuable than instrumentation.

6. Summary. The following steps should be followed in measuring accuracy:

a. Prior to collecting error data, perform an analysis to determine the following:

(1) The criterion of acceptable accuracy.

(2) The kinds of errors that may be made and those that will be recorded.

(3) The effect of the error on the individual task and the overall system.

(4) Whether errors should be recorded manually or by instrumentation.

b. In analyzing error data, weigh the error by its effect on the task.

c. Where possible, use an operating procedure as a guide to the recognition of error.

d. Where possible, determine the number of tasks successfully accomplished and apply the s/n ratio.

e. In a team situation, measure the accuracy of the team output and the individual contributions of team members to that output.

Frequency of Occurrence

1. Definition. The test planner may also wish to determine how frequently the operator's responses occur. Frequency, which is occurrence as a function of some time interval, is simply the tabulation of personnel actions (or events occurring as a result of personnel actions) as a function of time or other events occurring during system operations. In the latter case one might, for example, record frequency of error as a function of several stages of an operation. Frequency is easy to secure, provided one can sort personnel actions on a time base.

2. Information Provided. Occasionally, a standard of personnel performance will specify a required frequency of response (in which case it is necessary to verify that personnel do respond with that frequency), but this does not happen too often. Frequency data are of primary use in a supporting role; that is, to help explain other data more directly descriptive of personnel performance.

This information may illuminate factors that have affected the operator's performance. The relative frequency of certain types of error, for example, may suggest special difficulties the operator has in using his equipment. For example, if he makes a disproportionate number of errors in operating a particular control, that control may have been poorly designed or it may have been located in an awkward position. In evaluating the efficiency of a squad, the frequency of certain types of verbal reports from one squad member to another may indicate the degree of squad coordination.

3. Use Factors. Any series of discrete operator actions can be recorded as a frequency. Ths most common such applications are:

a. Types of personnel actions performed.

b. Types of error made in operating an equipment.

c. Types of verbal reports communicated from one action station to another.

d. Types of maintenance actions.

4. Problems. As with any other type of data, the actions whose frequency is being tabulated must be observable and distinguishable from other actions. Given this, it is relatively easy to secure frequency data. It is, however, much harder to interpret its significance, since frequency has only one dimension (occurrence) and does not by itself explain anything.

5. Example. Since many different functions (e.g., monitoring CRTs, communicating, analyzing map and status boards, etc.) are performed in TACC, it would be of interest to determine the distribution of these activities over time (an 8-hour shift, perhaps). Such data might have diagnostic value in terms of indicating where personnel are overloaded. For example, if at certain times one operator is burdened with a high frequency of telephone messages, it might be desirable to supply him with a backup communicator. Since, however, there is usually no system requirement that enjoins a particular frequency of activity, frequency data would be collected only if it did not interfere with other data collection requirements.

Frequency data can be secured in TACC at a number of levels. The most molecular would be the frequency with which a situation display operator uses certain controls in monitoring his assigned airspace (e.g., hooking, shifting sector scan, etc.). It is difficult to see the special utility of collecting this type of information, however, because, at this level, control use frequency is determined by the number and type of aircraft entering the airspace. Moreover, unless the data collection procedure was automated, it would tie up one data collector for each situation display operator being monitored.

At a higher level, the performance frequency of the functions (described at the start of this example) could be determined. This could produce useful if not critical information. Frequency at this level would also require direct observation by data collectors, although it would probably not require a data collector for each operator. This type of frequency information would be facilitated by use of a checksheet, enumerating the various activities to be performed; the data collector could simply check off each activity as he observed it occurring. Naturally, before developing this checklist, the test planner would have to analyze the types of activities that could occur. This analysis also presumes the observer's ability to distinguish one activity from another. If the activities did not appear with high frequency and did not overlap, the observer could also note the time of their occurrence and/or their duration.

The types of errors made could also be noted as they occurred. Assuming these do not occur too frequently, they could be noted as part of the rating evaluation of accuracy noted in the previous section. Frequency of error type has considerable diagnostic value, since error type is logically related to some problem in aptitude, training, procedures, or input load.

Determination of types of verbal reports would best be determined from analysis of automatically recorded communications, since verbal communication is usually quite rapid and the data collector may not have time to categorize the message. Communication frequency has some diagnostic value, but not as great as error type. Communications analysis has value in terms of suggesting the efficiency with which the team as a whole interacts among its members and with other organizations.

6. Summary. The following steps should be performed before measuring frequency of occurrence:

a. Define the performances whose frequency is to be measured.

b. Determine the time base (e.g., hours/minutes, events, stage of operations) to which frequency is to be related.

Amount Achieved/Amount Consumed

1. Definition. Some military systems may require that a specified quantity of a product be output by the operator and/or system. In other cases, the test planner may wish to determine what the operator/system is capable of outputting, even though no specific criterion requirement exists. The measure utilized for this purpose can be considered the amount achieved.

The mirror image of amount achieved is consumption or the quantity used, either absolutely or per unit of time; something must be used up which is related to the goal of system operation. The criterion (either explicitly or implicitly) is to use as few resources as possible. On a system level (e.g., artillery battery or tank), the number of shells expended, the number of missiles fired, the amount of fuel burned, etc. are sample measures of consumption.

2. Use Factors. In order to apply amount achieved, the operator/system must output a product which can be mathematically treated. The most obvious civilian example of such a product is the assembly line, where number of units being assembled, installed, welded, etc. can be counted to secure a total value. Military systems, too, have products that can be counted. For example, the number of shells being moved by a transportation company can be reported as an amount achieved.

One measure of amount achieved was referred to in the section on Accuracy. This is the percentage of tasks successfully acccomplished (however one defines successful accomplishment). The measure is simply the ratio between tasks successfully accomplished (s) and all tasks attempted (n). Even for those systems in which there is no overt product, one can consider the accomplishment of a task as a product which can be quantified by the s/n ratio. If one can observe the operator performing and say that a task has or has not been successfully accomplished, one can apply this measure. However, the tasks must be discrete and there must be a concrete criterion of successful completion. This criterion need not itself be quantitative, as long as the evaluator can judgmentally determine task accomplishment.

The s/n ratio is, however, purely a binary measure and does not allow for degrees of success; in that respect, it is a rather crude measure. Moreover, to apply the ratio, any task being evaluated must be repeated a sufficient number of times. If the operator performs a particular task only once or twice, he may either succeed completely (thus giving him a performance value of 1.0) or fail totally (s/n ratio of .00). Either of these values is obviously a gross over- or underestimation of his true performance capability.

If the system has a specific quantifiable goal (e.g., hitting target X ± 100 yards), then a deviation measure (the extent to which rounds exceed allowable tolerance) can be utilized as a measure of accomplishment. The number of rounds on target or the time on target are also measures of amount achieved.

In most cases, system requirements will not specify the amount (of any product) to be achieved. Hence, the usefulness of this measure is primarily in describing how the operator/system performs and not to verify a given level of performance.

3. Problems. It is not easy to measure personnel performance achievement, because the product of personnel performance is often not concrete, as for example, in monitoring, planning, or deciding. Any overt personnel response can be counted; but even so, such a measure might have little relevance to the goal of the total task. For example, one could count the number of switches thrown by a console operator, but this number might have little bearing on what the operator is attempting to do, which might be to vector aircraft. From a personnel performance standpoint, consumption also is not an easy measure to apply since the operator often expends nothing of his own except his energy.

If a team has a concrete output, it can be measured in terms of accomplishment just as the individual operator's output. For example, consider an air defense center with six operators, each of whom is responsible in his own sector for detecting, tracking, and assigning weapons against enemy fighters and bombers. Each operator's performance can be measured in terms of the number of targets assigned to him and the number he destroys, or the number he permits to reach some minimum distance from the center. Likewise, the number of aircraft or missiles he assigns against these targets can be considered a consumption measure.

To measure the center's performance, the performances of the individual operators can be combined; for example, if each of the 6 operators had 12 targets to handle (a center total of 72) and the total targets destroyed were 63, then the center's performance would be assessed as 63/72 or 87 percent effective.

4. Example. Amount achieved is not an easy measure to apply in the TACC context because, strictly speaking, there is no product of TACC operations. It is, however, possible to apply the s/n ratio to individual functions such as air intercepts with individual aircraft (because these are essentially discrete operations). Other TACC activities (such as communicating or coordinating) cannot be evaluated objectively as successful or unsuccessful.

A consumption measure can also be applied to the number of aircraft launched by TACC to contain an aggressor force or the number of weapons it assigns to counter a threat. Such a consumption measure, however, applies only to a subset of TACC activities (countering a threat) and can be used to describe the cost of that subset alone.

5. Summary. The following steps should be followed in deriving a measure of amount achieved and/or consumed.

a. Determine what the operator's responses are.

b. Determine that they are discrete and quantitative (or capable of being scored quantitatively).

c. Determine that they are meaningfully related to the overall goal of system performance.

d. Specify a number that expresses that relationship.

Subjective Measures and Methods

Discussion

Superficially, the ideal measure is objective and quantitative. Although objective measures are necessary, they are insufficient because the very few dimensions they possess (e.g., time and amount) do not permit them to describe performance fully. Simply to know that operators have achieved a certain level of accuracy or take a given length of time to perform a task does not tell us all we wish to know. This is particularly true if operator performance exhibits deficiencies; the evaluator will wish to know what causes these and a single number or even several numbers will not usually supply this information.

In addition, the performance one wishes to measure may be difficult to specify or secure in objective terms. Suppose, for example, that one of the criteria to be applied in evaluating a new reconnaissance vehicle is comfort in riding the vehicle. It is unlikely that one could measure this quality solely by objective means.

Under these conditions the evaluator must apply subjective methods. One can in fact lay down a principle: Any evaluation will be incomplete unless subjective data are gathered in addition to those of an objective nature.

Subjective data are secured primarily through human judgment, but the demarcation between objective and subjective methods is sometimes difficult to specify. For example, if a data collector usually counts the number of holes in a target on the firing range, is this objective or subjective? The measure is secured by perception, but the act of counting is ordinarily so unequivocal that one would hesitate to apply the term subjective to it.

Objective measures must be quantitative ones; subjective measures may be quantitative or qualitative, depending upon how they are treated. Subjective measures are not automatically qualitative, since there are many ways of quantifying them (e.g., see later section on ratings).

A more serious shortcoming of subjective measures is the fact that they are more inaccurate than objective measures since the former depend more on human observation, which is often imprecise. Inconsistency (lack of reliability) is another defect; the individual who supplies subjective data may be inconsistent from one measurement to another, and of course individual observers vary among themselves. This variability may be greater or less, depending upon what is being observed. Much subjective variability results from the fuzziness of subjective criteria; for example, if the performance criterion for a squad is that it should be "highly coordinated," what is one to look for to arrive at this conclusion? If subjective criteria could be specified precisely, subjective data would be as precise as objective data.

One cannot, however, avoid using subjective methods for these reasons:

1. They provide data that cannot be supplied by more objective methods. For example, it is essential to the evaluator to determine how

the personnel whose performance is being evaluated feel about the test conditions and their own state relative to test inputs (e.g., fatigue, motivation, difficulty of the situation, problems experienced, etc.). That is because how the operator views the test situation may well determine how he performs in that situation.

2. Some data can be secured only through subjective methods. For example, it may not be clear exactly what has taken place during the operator's performance; questioning the operator can help clarify this. The factors impacting on the operator's performance may be invisible except to him.

3. Subjective methods are often less expensive to employ than objective ones. Where objective measures require expensive instrumentation which is difficult to operate and maintain (in a field environment, for example), it may be necessary to seek the same information in a different manner. There are, however, personnel costs associated with the use of subjective techniques. Subjective methods can be used to secure data on the following:

a. Description of performance (what took place).

b. Determination of performance effectiveness (by means of ratings).

c. Factors subjectively experienced by the operator that affect his performance (e.g., fatigue).

d. Adequacy of system characteristics such as the procedures used.

e. The operator's internal state (e.g., motivation).

The Interview

1. Definition. The interview is one of the most common methods of securing subjective data. In essence, it is simply asking questions of a respondent orally and noting his answers. The interview is difficult to apply well because it is the least structured subjective method (often taking form as the interview proceeds) and because it involves interpersonal relationships between the performer and the interviewer (e.g., the performer's reactions to the interviewer's manner).

Interviews may take various forms:

a. Individual (one interviewer, one performer). This is the most common form of the interview.

b. Team (one interviewer, several performers being interviewed as a group). This occurs only in connection with team operations.

c. The interview may be combined with other subjective methods such as rating scales, which are completed by the interviewee.

d. The interview may be combined with a demonstration of performance by the operator.

2. Use Factors. An interview should always be held at the completion of the test; no performance evaluation should be considered complete without it. However, the interview cannot substitute for more objective methods. It cannot, for example, be used to verify that performance satisfies system requirements or to predict future operational performance. However, it can help to diagnose problems encountered in testing and to identify required personnel resources and training.

Since the interview is directed at securing additional material concerning prior performance, the interviewer should also have observed that performance.

What types of information would one wish to secure from the interview? Literally anything about which one wishes an opinion from the performer can be secured from the interview, but the following questions are the most common. (Note: These are not necessarily phrased in the form in which they would actually be asked.)

a. What did the performer do while he was performing? For example, when the operator field stripped his machine gun, did he remove the magazine before he broke it down? Unless a record of the subject's performance was made automatically (e.g., by videotape camera), the interview may help to check the accuracy of the data collector's observations.

b. Why did the subject perform as he did? For example, why did the squad leader select one reconnaissance route rather than another? If the performer made errors, why did he?

c. What knowledge does the performer have about the principles and information that should have guided his task performance?

d. What test conditions (e.g., night/day) affected the performer most and why?

e. How well did the operator think he performed? If there are any significant discrepancies between the subject's performance as it was observed and his own evaluation of that performance, why?

f. With what tasks did the operator experience the most difficulty? Why? What factors does he feel contributed to those difficulties?

g. In a team operation, how was responsibility divided among the members?

h. Citing a specific factor of interest to the evaluator (e.g., reduced visibility, logistics, etc.), what effect did that factor have on the operator's performance?

i. Does the subject have any comments at all that he wishes to make about any aspect of the test or his performance? (This is an open-ended question, usually asked at the end of the interview, and serves as a prelude to closing the interview.)

The interview should be partially structured and partially free to vary as the subject's responses suggest. Those key questions for which the investigator thinks he needs answers should be developed in advance (although they need not be written out). A key question can be used to initiate the interview. The subject should be allowed to expand on topics of interest to him but, when the interview tends to wander, the interviewer should bring its direction back to the point by asking another of the key questions. If the subject appears reticent, the interviewer can use his prepared questions to stimulate his responses.

Where the performance being evaluated is that of a team, the interview should be conducted with the team as a group. If the team consists of many members, it may be unfeasible to interview each one individually and, in any event, individual interviews do not permit the evaluator to record team interactions.

Since performance about which the operator is being evaluated is technical in nature, it is essential that the interviewer know at least the gross technical details of the task, although he need not be as proficient as the performer in that task.

The interview is commonly held after the task has been performed or at some convenient interval in system operations. Where a series of tasks is being performed, the interview should not, if possible, be postponed until every task is completed, because then the subject's memory for events is reduced. It may not be feasible, however, to hold an interview immediately following the performance of each individual task. Obviously, one cannot break in on an on-going system operation to interview; consequently, the interviewer should look for "natural" breaks in the flow of system events to ask his questions.

The subject should be interviewed concerning each task he has not performed previously. If he repeats the same task on several occasions, and task conditions remain the same, it may not be necessary to interview repeatedly. However, it is advisable to secure two interviews (separated by an interval) for each task performed by the operator to determine the consistency of his interview responses. If test conditions do not change substantially on repeated performance of the same task, the answers the interviewer gets on the first two interviews will probably be repeated in subsequent interviews, and the performers will find the questions boring. Under these circumstances, the scope of the interview can be reduced. If there is some expectation that subject responses will change in subsequent task performance (because of learning, for example), the subject should be reinterviewed, although the length of the interview can be shortened when it is repeated.

The length of the interview should depend on how much the performer can tell the interviewer, but 20 minutes is a good average length. Beyond 30 minutes, the subject tends to become fatigued.

The conditions under which the interview is conducted are important. Ideally, it should take place away from the test operation (although not too far away lest one has to account for transportation time), in a fairly quiet

place, with no interruptions or distractions. A specific room for the interview would be most desirable since this helps to emphasize the importance of the interview to the performer. The fact that an interview will take place should be made known to the subject in advance. He should be reassured at the outset that nothing he says will be held against him, and that he will not be identified in records or reports. He may ask for feedback about his performance since he is understandably concerned about his proficiency in an evaluation. He should be told that the interviewer is not the evaluator but that it appears to the interviewer that the subject's performance was certainly adequate.

The most convenient way of conducting an interview is to tape record it but, if it is a very short one, or the interviewer is highly skilled, it can perhaps be handled by note taking. As far as possible, interview responses should be recorded verbatim without the interviewer making any attempt during the interview to analyze them.

4. Problems. One major advantage of the interview is that it is an interpersonal situation. It therefore can be used to motivate the performer to communicate by suggesting that his performance is considered important. Beyond that, many people feel less constrained in responding orally than they do in writing. Another advantage is that the flexibility of the interview permits the subject to partially control it by selecting or emphasizing topics as he wishes.

One of the disadvantages of the interview is that it is usually a one-on-one procedure, which means that a great deal of time can be consumed with this method. Since conversation is not very structured or standardized, the analysis of the material gathered is more difficult because the freedom of the procedure permits the interviewee to include irrelevant and unimportant responses.

5. Example. Interviews in the TACC evaluation situation could be either individual (one interviewer/one respondent) or group (one or more interviewers/the TACC team or some major part of it). Because TACC is a team operation, it would be reasonable to interview the team as a group. Alternately, since TACC personnel perform individual (although overlapping) functions, it would be possible to interview at least the key personnel of the team individually.

The most reasonable time for the interviews was at the end of the work shift, because it was impracticable to break into a "loose" operation like TACC that has no natural break points. The interviews, conducted by those who had been observing the previous work shift operations, were held in an administrative area located next to the TACC shelter.

It was eventually decided to hold an initial group debriefing session after the first work shift, to be followed by relatively short individual interviews with key personnel (excluding auxiliary personnel such as map and status board operators).

Interviews focussed on determining where difficulties had arisen or where problems might arise (e.g., points of heavy target inputs). The group

interview began with a recapitulation of the major events as these had occurred during the work shift. As each activity was reviewed, the individuals involved were invited to comment on the accuracy of the recapitulation and to supply additional information concerning these events. Because this was a team interview, special attention was paid to coordination within TACC and between TACC personnel and other organizations. The TACC team was invited to report any difficulties they experienced and to suggest improvements that could be made in operating procedures, technical data, etc. The interviewer took notes but the proceedings as a whole were tape recorded.

Subsequent individual interviews were held with key personnel. These dealt in greater detail with the specifics of their individual work stations and functions. Because the group interview had preceded the individual ones and had, presumably, elicited major problems, the individual interviews were substantially abbreviated.

The TACC evaluation ran daily over a period of approximately 6 weeks. During this time, a group/individual interview was held at the beginning, midway through, and at the conclusion of testing. The Chief Evaluator felt that such a schedule was not an inordinate imposition on TACC personnel and was actually the minimum he should have to gather interview data.

6. Summary. The following steps should be performed before conducting the interview:

a. Determine the information you wish to secure from the performer.

b. Develop key questions to elicit that information.

c. Within bounds, allow the interview to "go" the way the respondent wishes it to go.

d. Observe the task performance for which you are interviewing.

e. Set a maximum interview length of 30 minutes.

The Questionnaire

1. Definition. The questionnaire is a more structured form of the interview, structured because the questions are written and cannot be modified for the individual respondents.

2. Information Provided. The uses of the questionnaire are the same as those of the interview: diagnosis of problems encountered in testing and the identification of required personnel resources and training.

3. Use Factors. The questionnaire has certain advantages over the interview. Because the questions are written, the questionnaire can be administered to many individuals at one time, which makes it more economical. (Obviously, the questionnaire is particularly suitable for teams.) Since the questions asked are identical for each respondent, the variability between interviewers (which is inherent in the interview) is avoided. Because all subjects respond to identical questions, it is easier to compare responses

made by individuals. The consistency with which the individual answers can be tested by asking the same question in different forms. Although written materials (e.g., rating scales) can be administered as part of the interview procedure, it is usually easier and more convenient to administer these additional materials as part of a questionnaire.

The structure in the questionnaire also makes it somewhat easier to analyze its data. The questionnaire can be more readily tried out and its items modified to improve their precision than can the same questions in interview form.

The questionnaire has disadvantages, however. Since the questions asked are presented in a fixed order and with fixed content, the subject has less freedom in responding. If he should wish to comment on some topic not included in the questions, he cannot do so, although it is possible to pick some of this material up by providing open-ended questions asking for such comments. The subject may not understand a question and has less freedom to ask for an interpretation, although the questionnaire administrator is often available to provide guidance. The space permitted individuals in which to write is necessarily limited; many people who do not like to write will be less responsive than if interviewed. Because of the impersonality of the questionnaire format, subjects may be less motivated to respond fully. Respondents often pay less attention to and devote less care to completing a written form. The questionnaire administrator cannot ask for explanatory material when he desires it.

For all these reasons, the questionnaire is less preferred than the interview as a means of gathering data. However, it can be used to collect the same types of information as the interview (for a list of these, see the preceding material on interviewing).

4. Problems. In both the interview and the questionnaire, but more so in the latter, detailed instructions on how to respond must be provided to the subjects. Information about who has developed the questionnaire and for what reason should be provided. Questionnaire administration requires a larger office area because one is almost always dealing with a group. Writing surfaces and pencils must be provided.

Like the interview, the questionnaire is also administered after test performance. However, the events to be reported on are more likely to cover a longer period of test time, because it is unfeasible to convene a group too frequently. When questionnaires are very lengthy or where subject time is very precious, they may be administered on a "take home" basis, with the subject responding when he has time; however, this procedure is not desirable because the manner in which the respondent completes the form is uncontrolled. Because of the general reluctance to write at length, the questionnaire should be as short as possible while still soliciting the desired information.

All other provisions that apply to the interview apply also to the questionnaire.

5. Example. Because of the number of TACC personnel, which made interviewing expensive, evaluators relied heavily on the questionnaire as a source of data. On a predetermined schedule (at least once every three shifts) questionnaires (requiring, on the basis of tryout, not more than 15 minutes to complete) were handed out and personnel were asked questions such as:

a. How would you rate your performance on this shift compared to that of the preceding two shifts?

b. Were there any times during the shift when you felt you could not keep up with the input flow? (If so, describe.)

These questionnaires were reviewed and, if any response appeared to warrant further explication, interviewers contacted the respondents. The questionnaires supplemented the observational data collected during the test by allowing subject personnel to report data to which they alone had access.

6. Summary. The following steps should be followed in applying this method:

a. Use the questionnaire when it is necessary to collect data from many subjects and/or sufficient interview time is not available.

b. Reduce the length of the questionnaire as much as possible.

c. Try out the questionnaire, analyze the try-out responses, and interview respondents concerning their understanding of the questionnaire items.

d. Provide detailed instructions on how to complete the questionnaire.

e. Wherever possible, administer the questionnaire in the presence of an evaluator who can answer respondents' questions.

Observation

1. Definition. Observation is the collection of data primarily through perception and analysis of the event being observed. The emphasis here is on the adverb "primarily," because all of the methods discussed so far involve at least a minimal degree of observation of, for example, data collection instruments. Observation in the sense in which we refer to it is closely associated with ratings (to be discussed later).

2. Information Provided. The importance of observation is illustrated by the fact that it alone can supply data for all the purposes of an OST (see Table 3-1). More specifically, it can be used to:

a. Describe what events/responses occurred (what took place, who did it, when and with whom did it occur)? In highly proceduralized performance situations, this use of observation may be minimal.

b. Determine the frequency of system events or operator responses.

c. Determine the accuracy (or errors) with which these events/responses took place.

d. Infer certain qualities in performance and make judgments about that performance, as in the case of rating the efficiency of performance.

3. Use Factors. Observation is not the only method available to supply these data. Except for ratings, automated methods can be applied in each case (e.g., videotape recording) to describe what took place. Observation is merely one of the evaluator's options. Moreover, it can be used with instrumentation. For example, a videotape camera can be used as a back-up to direct observation when inputs are so numerous or so rapid that the observer cannot deal with all of them.

Observation is therefore an alternative to automatic measurement. Where instrumentation is not available or feasible (from an engineering or cost standpoint), observation becomes an essential means of gathering data. For example, when a system is evaluated during field exercises, observation may become a very desirable alternative to instrumentation.

4. Problems. Observation appears to be so "natural" a method that the test planner may tend to overlook the problems it brings. One problem referred to in previous discussions is the definition of what constitutes the system events or personnel responses to be observed. The written description of a task often varies in some dimensions (e.g., flow of inputs, team member interactions) from the task observed during performance. Written language is unable to describe fully the richness of the task as performed, especially when the task is complex (e.g., continuous or involving a team) or consists largely of covert behavior. The observer must be able to recognize (a) when the task starts and when it is completed (relatively simple), (b) which operator responses are correct and incorrect and why (somewhat more difficult), and (c) which task qualities are relevant to the purpose of the evaluation (much more difficult). For example, to determine that fatigue is affecting performance is often extremely difficult, because fatigue indices are usually not obvious.

Extensive experience in performing the task to be observed is highly desirable, but not necessary, although the observer lacking any experience must be given some training in how to perform the task. Even when he is experienced, the observer's experience of the task as he has seen it performed may not correspond entirely with the task as performed during the evaluation. For this reason, training in observation is essential.

Observation is not merely perception ("seeing"); it also involves analysis and interpretation, by the selection of relevant task characteristics. Often this analysis/interpretation is completely unconscious, but training can make the observer more aware of it. One goal of observer training is to make relevant task dimensions more visible to him.

The observer himself is the limiting factor in observational accuracy. If many task events occur rapidly or concurrently, the observer may not pick them all up. He sees, but he may not be aware of what he sees. The well known unreliability of witnesses to criminal actions is a case in point. When the observer is time or event-stressed he may:

a. Select a subset of task events to report, ignoring the remainder. What he reports may therefore be incomplete and only partially relevant.

b. Use gross rather than detailed reporting categories.

These difficulties are particularly evident when the observer rates the operator's performance. Ratings are apt to be based on several dimensions (e.g., the operator's speed, accuracy, coordination, etc.). If the observer rates a task performance as "adequate," what does this mean? Gross categories of performance, such as adequate, acceptable, satisfactory, etc., usually involve several elements, some of which the observer may ignore in making his judgment. Did the observer actually note these elements and did he weight them correctly when he rated the performance?

As the complexity of the task being observed increases, it becomes more desirable to have two observers whose results can be correlated for consistency. However, this increases the number of required data collectors.

Observers must, therefore, be familiar with the tasks to be evaluated and they must have an opportunity to practice observing and reporting their observations. Observations made during training should correspond as closely as possible to the conditions under which actual observations will be made. If possible, criteria of observational accuracy should be established and used during training to evaluate the observer.

Training also presupposes that instructional materials have been provided the observer. These should include the following:

a. Specification of the task characteristics the observer should be looking for, together with operational definitions of these characteristics.

b. A properly designed form with which to report the events observed.

How the observer reports his observations may significantly affect his accuracy. The following are possible observational formats (not all of them are equally desirable):

a. The observer may attempt to write down or tape record everything he has observed or heard. This is undesirable because he will probably leave out certain events.

b. If the task is highly proceduralized, he may use an operating procedure as a guide to task performance and check off each step as the operator

makes it (see, for example, Figure 3-3). However, this may be difficult if the task is continuous or the sequence of operator responses is variable (contingent).

c. He may develop a checklist of significant aspects of the task (those actions most likely to be performed erroneously or those that may be most difficult for the operator) and use this as a guide to what he should observe.

d. He may combine formats b and c, using an operating procedure as a general guide and noting all errors or other behavior of interest as these occur.

The format selected will depend at least in part on the nature of the task to be observed.

If the observer is physically to see enough, he must be close to the task being performed. If he is too close, however, he may intrude upon the performance and thus distort the way in which the task is performed. Every observer must be considered as potentially obtrusive, even when the operator knows why he is being observed and is familiar with the observer.

5. Example. In the TACC evaluation, observations were made as part of the process of rating individuals and the total team on their efficiency in monitoring airspace, launching interceptors, coordinating assault forces, and interacting with other team members and other air defense organizations. Only the 11 key personnel were observed and observations were made only on a sampling basis. Prior to the start of the evaluation, test planners met at intervals over several weeks to define as precisely as possible what was to be observed. Questions raised at these meetings were:

a. Is the desired behavior reliably observable?

b. Is it meaningful in terms of the evaluation purposes?

c. What forms will be used to record the data?

d. How will the observation/ratings be analyzed?

Rating forms to be discussed subsequently were developed to include these definitions. The first three or four observations were considered as essentially practice sessions and the results achieved with these were analyzed by the test planners for consistency, difficulty in making the observation, and accuracy of the results. Consequently, the observational methods with which the study began were somewhat revised before planners were satisfied that valid data could be secured.

6. Summary. In making use of the observational method, the following steps should be followed:

a. Define as precisely as possible what is to be observed.

b. Select subject matter experts as observers (because these are easier to train, having had prior experience) and provide systematic observer training.

c. Develop a reporting form that provides cues as to what is to be seen and that indicates as precisely as possible the categories with which the observation is to be reported.

d. Measure the consistency of individual observers on repeated observations and the consistency between observers. Reject those observers who are not consistent in their performance.

Ratings

1. Definition. Ratings are the means by which observational judgments are quantified. There are several kinds: (a) ratings by a data collector/evaluator, (b) supervisor ratings (where a team is involved, one of whose members is in charge of team activity), (c) peer ratings or ratings by one member of the performance of other team members with whom he has interacted, and (d) self report ratings in which the operator judges his own performance or his own feelings regarding an event or performance that he has accomplished.

The difference among these ratings is the rater's involvement in the activity being rated. In self ratings, he evaluates his own performance; in peer and supervisor ratings, he is only a partial participant in the activity being evaluated, and in data collection/evaluation ratings, he is a nonparticipant. Most ratings in the OST context are data collector ratings, but there is also value in the other ratings. Conceivably, the more involved the rater is, the more valid his ratings, but one must also consider bias.

Ratings are quite popular in personnel performance measurement for two reasons: (a) assigning a number to a subjective assessment makes it possible to treat that assessment quantitatively, and (b) most rating scales used for operational evaluation can be developed quickly, are easily used, and require little or no specialized training.

Having explained rating popularity in this way, it is necessary to say also that applying numbers to a subjective assessment does not make that measurement more accurate or valid. To utilize ratings correctly does, in fact, require extensive development, validation, and training of the rater.

This discussion does not pretend to be a highly technical analysis of ratings and rating scales, about which much theoretical material has been written; only the essential points relevant to OST are covered here. The reader interested in more detail should see Guilford (1973).[2]

---

[2]Guilford, J. P. Psychometric Methods. New York: McGraw Hill, 1973.

2. Information Provided. When a judgment is made that a particular performance is satisfactory or unsatisfactory, ratings can be used to verify performance. To the extent that verification represents how the system will do operationally, ratings can be used to predict operational performance. The same data can be applied to future systems. When a performance rating reflects on the training required to achieve that performance rating, it can also be used to provide data for future training.

Literally anything can be rated, assuming one can scale it. Ratings should be used, however, only when an objective method cannot be found, unless cost or feasibility considerations make the latter undesirable. Ratings are most often employed for the following purposes:

a. To evaluate how well someone is performing a task (this rating can be made by an observer or by the performer himself).

b. To evaluate some quality of task performance, (e.g., coordination of squad members).

c. To evaluate the adequacy of some feature of the system being evaluated (e.g., displays, procedures, and logistics).

d. To evaluate the effect of some condition (e.g., visibility) which impacts upon performance (e.g., detection of targets).

e. To evaluate the output of task performance (e.g., the adequacy of a tactical decision).

3. Use Factors. Since ratings depend on observation, rating validity (the extent to which the rating mirrors real phenomena) depends on the observer as a measuring instrument. Everything said in the previous section on observation therefore applies to ratings.

There are several types of rating scales, but since the graphic scale is the most popular and the most adaptable to OST, this is the one that will be described.

The graphic rating scale can be designed in different ways by taking a straight line (representing the scale along which the variable varies) and combining it with descriptors that represent the values of the variable being scaled. The line can be segmented in units or it can be continuous. If segmented, the number of segments can be varied. The line can be placed horizontally or vertically. The number of scale points can be varied.

A number of guidelines for development of the scale have been laid down:

a. The line forming the scale should be 5-6 inches long.

b. Continuous rather than segmented lines are preferred because the former suggests that the variable being measured is continuous.

c. Where several scales are used to measure individual dimensions of the variable, the ends of the scale representing the "good" or "high" aspects should all be in the same direction.

d. If the scale is vertical, put the good or high end of the scale at the top of the page.

e. The descriptors of the various points along the scale need not be equally spaced along the line, although this is general practice, because usually one does not know how the variable being rated deviates from linearity.

f. The descriptors are placed opposite the scale points which they describe.

g. Descriptors whose meanings or values are so extreme (e.g., never or always) that they will almost never be applied by the rater should not be used.

h. When the variable being evaluated is bipolar (i.e., can be described either positively or negatively, as, for example, speed ranges from slow to fast), the neutral point is midway along the line.

i. The number of scale points generally vary from 5 to 9. Fewer than these leads to grossness and lack of precision in the rating; more than these cannot be meaningfully discriminated by the rater. The precise number to be used depends on the nature of the variable and the evaluator's ability to think of distinctly different steps along the scale for that variable. A 5-point rating scale is most common, however.

Graphic rating scales are fairly simple to develop and easily administered. This says nothing at all for their validity or the process of validating them, which can involve considerable effort and which is why most scales used in OST are not validated. If used for self rating performance, they are interesting to the rater. They are easily completed, since the rater does not have to perform any mental calculations. They provide for as fine a discrimination as the rater is capable of and as fine a degree of scoring as the evaluator desires.

The first (and most difficult) steps in the development of the rating scale are (a) to conceptualize the performance variable (e.g., speed or degree of comfort experienced) to be rated in terms of a continuum, and (b) to segment that continuum into psychologically equal intervals.

The performance attribute to be rated must be observable by the observer or the task performer, which is simply another way of saying that whatever is to be rated should be capable of being rated. If scales are developed for an attribute which the rater cannot validly observe, the resultant ratings will be invalid, even though numbers have been applied.

Everything said in Section Two concerning criteria and measures in the Personnel Performance Test Plan applies also to the development of ratings. Before the rating scale is developed, it is necessary to operationally define the performance variable to be rated. It is easy to scale any performance in terms of values such as excellent, very good, good, adequate, slightly inadequate, poor, and terrible (a seven-category scale), but what do these phrases mean? Suppose one wished to rate ability to perform ground reconnaissance, using the previous categories. What would a rating of adequate in this situation depend on? Speed of detecting certain terrain or man-made features? Speed of reporting information? The amount of detailed information provided in the report? And if so, how much speed or information denotes adequacy? Obviously we come immediately up against the problem of criteria. Where objective criteria are not available, they can be developed using the techniques described in Section Two. Because the criterion must be scaled to make it suitable for rating, it is necessary to fractionate it into equivalent parts, which means that a criterion must be developed for excellent, very good, good, etc.

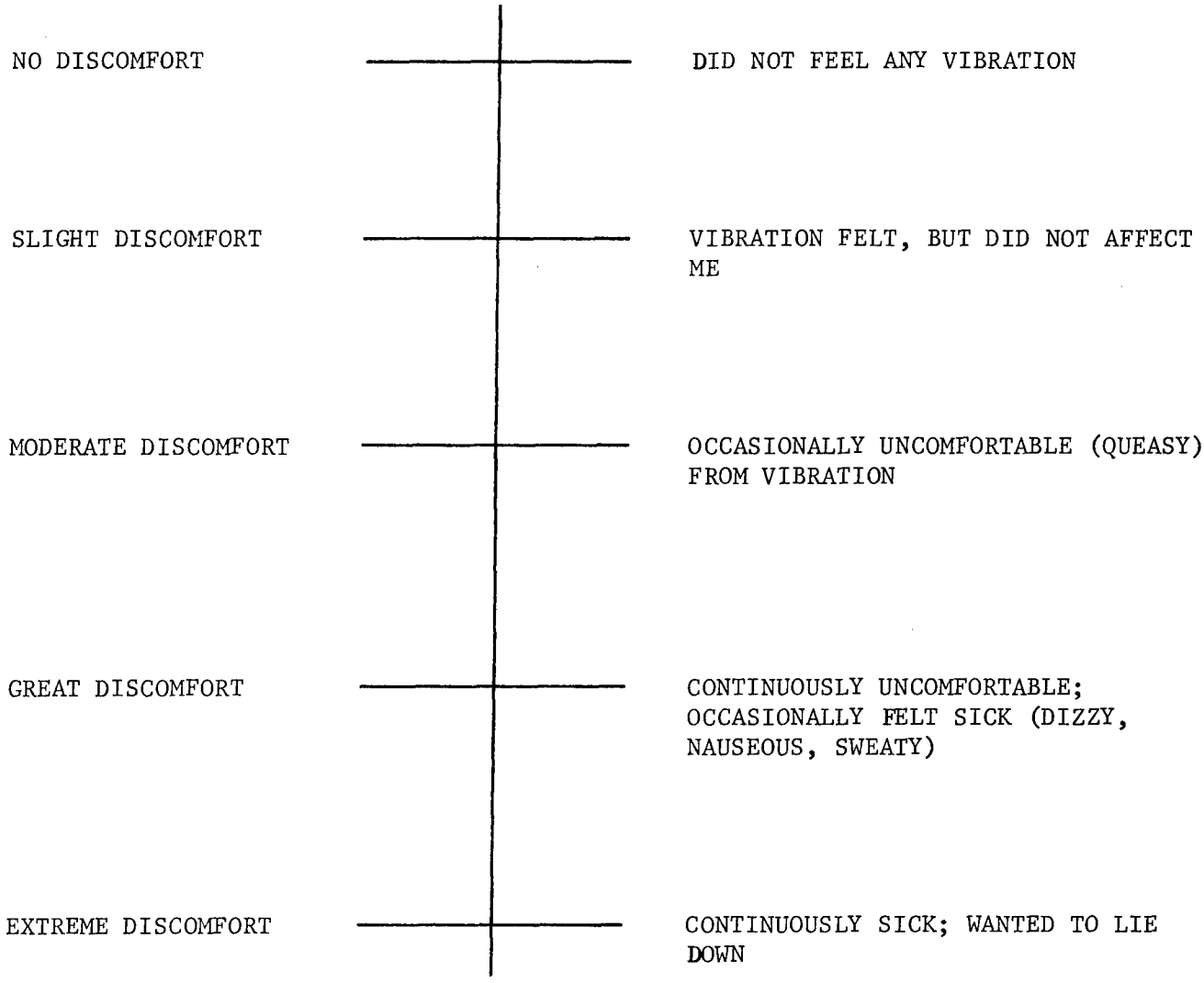
The rating is particularly useful when performance depends on multiple correlated dimensions (such as adequacy of ground reconnaissance). It may not be feasible to develop individual, independent measures of each dimension and the integrated criterion may not be susceptible to instrumented measurement. In this case a global rating is appropriate provided the rater is aware of the dimensions making up this rating.

As an illustration of the preceding concepts, assume that the variable being rated is the amount of vibration experienced by troops being transported in a LVTP-7. Note that this is not the same as the degree of vibration physically occurring in the carrier; the latter can and should be more effectively measured by means of accelerometers. What is being rated in the sample scale (Figure 3-4) is the individual Marine's experience of that vibration (this is a self-report scale), which cannot be physically measured. This is a good example of a subjective measuring tool which complements an objective one. Both objective and subjective measures are desirable, since the experienced vibration and its effects may not be linearly related to physical vibrations.

The first thing to do is to decide on the dimensions of the vibrations (i.e., the basis for making judgments) as conceptualized in Figure 3-4. There are two dimensions in the Marine reaction to vibration: (a) the frequency of physical symptoms resulting from the vibration (e.g., never, occasional, continuous), and (b) the severity of those symptoms, (e.g, discomfort, sickness). These dimensions can be varied interactively on a scale of discomfort. (Note that, since vibration can have only negative effects, this is not a bipolar scale, so one need not worry about the neutral point between the two poles.)

Once these dimensions are conceptualized, the extremes of the scale values are developed: no discomfort and extreme discomfort. Intermediate values can then easily be supplied: slight, moderate, and great discomfort. This gives us a five-point scale.

Since discomfort is represented by more than one dimension, it is necessary to supply more detailed descriptors at each scale point, as shown in Figure 3-4.



Definition: Vibration discomfort is any feeling of unpleasant physical sensation experienced by personnel as a direct result of vehicle vibration.

Figure 3-4. Vibration discomfort rating scale.

Note that the two dimensions (frequency, extent of symptoms) interact, as in the scale item "great discomfort" where the Marine reports that he was continuously uncomfortable and occasionally sick. The respondent can select any point on the scale (including those between descriptors) that represents his feeling of discomfort.

A number of questions can be asked about such a rating scale. Do the two dimensions (frequency and severity of symptoms) truly represent the elements of discomfort as actually experienced by personnel? This can, perhaps, be determined in advance by questioning personnel who have experienced discomfort in transport vehicles.

A second question is whether the intervals between scale values are equal (which is the way the sample scale was devised). Is the difference between NO DISCOMFORT and SLIGHT DISCOMFORT the same as the difference between GREAT and EXTREME DISCOMFORT? This is a tough question to answer. If the range of discomfort is equally distributed in personnel in real life, one should find an approximately equal distribution of scale responses if one tested a large enough sample of respondents. However, in actual practice, one simply assumes equal intervals and lets it go at that.

The scale points, it will be noted, have not been numbered (to avoid biasing the subject), although this is entirely possible. In this case NO DISCOMFORT represents 0 (zero), EXTREME DISCOMFORT, 4. Since the respondent can mark anywhere along the scale, one need only measure the distance of the respondent's check mark from the nearest descriptor to secure a quantitative value for his response. For example, if he checks half way between SLIGHT (1) and MODERATE DISCOMFORT (2), he would have a score of 1.5 on this scale. Precision of scale measurement can be anything the evaluator desires, but in a relatively crude instrument like this one, he would probably not attempt to measure to more than tenths (e.g., 1.5, 1.6, 1.7, etc.). Precision is not the same as accuracy or validity.

An attribute to be rated should ordinarily be introduced with a name and a definition phrased in operational terms; these should provide cues to the rater to permit him to recognize the attribute. The attribute name without accompanying definition is likely to be misleading.

Rating scale descriptors such as "occasionally uncomfortable (queasy) from vibration" (see Figure 3-4) have the double function of supplementing the definition of the variable being rated and of providing anchors or milestones to guide the rater in making quantitative judgments. The following criteria should be followed in constructing scales:

a. Clarity. Use short, simple, unambiguous terms.

b. Relevance. The descriptor should obviously be consistent with the attribute being rated and its definition.

c. Precision. A good descriptor applies to a very short range on the variable continuum. There should be no doubt about its rank position relative to other descriptors.

d. Variety. Vary the language for descriptors at different scale points.

e. Objectivity. Descriptors that imply that a certain behavior is good or bad, desirable or undesirable, should be avoided.

f. Uniqueness. Descriptors should be unique to the attribute being rated. Where possible, avoid very general descriptors such as "excellent" or "adequate."

Experience has taught us a good deal about raters and their peculiarities:

a. Individuals differ in their capability to rate others.

b. Two ratings by the same rater are no better than one because the rater tends to repeat his errors a second time.

c. Raters need sufficient time to make their judgments.

d. Raters do better if they know the task whose performance they are evaluating.

e. Self-raters tend to overestimate more than underestimate their performance on the variables they are rating.

f. The assurance of the rater is important. Judgments of which he is very sure are much more reliable than other ratings.

g. Different raters often use different criteria in judging. This makes systematic training in rating all-important.

When finally developed, the rating scale should be tried out to see how well it performs. Although complex experimental and statistical procedures for scale validation exist, these are usually not feasible in an operational test situation (e.g., lack of time). Where possible, however, the prospective scale should be applied to two separate subject samples, known in advance to vary on the performance variable being rated, to see if the scale actually differentiates between them. For example, to rate the adequacy of maintenance on an equipment rating scale, evaluations of experienced technicians and students in training might be compared to see if the scale will provide differential answers (as it should, because one would logically expect these two types of personnel to behave differently in troubleshooting).

If this is not possible, the scale should at least be tried out with a number of raters to ensure that the meanings of terms are understandable, that there is no confusion between descriptors, and that, in general, problems will not be experienced in using the scale.

Since the measuring instrument in the rating scale is the rater himself, it is critical that he be trained in the use of the scale before he actually applies it. Training should involve the following:

a. Analysis of the operational definitions of the performance variable and the individual scale points.

b. Examination of the criteria inherent in the scale.

c. Practice in rating the performances to be evaluated.

d. Comparison of rater consistency with the results of others who are independently rating the performance. We deliberately have not mentioned rater accuracy, because this assumes an independent measure (which one assumes to be valid) of the variable being rated. Such independent measures are rarely available.

e. Training in the avoidance of common errors made in using rating scales.

4. Problems. Experience with ratings over the years has identified a number of errors commonly made by raters that lead to rating inaccuracies.

a. The error of leniency. Raters tend to rate those whom they know well or with whom they are ego-involved higher than they should. Some raters are "hard," others are "easy." Easy raters are more common. Any tendency to be unduly easy or hard in rating should be determined during training practice sessions and this tendency should be brought to the rater's attention.

b. The error of central tendency. Raters hesitate to make extreme judgments and thus tend to "bunch" their ratings in the center of the scale. This reduces the discriminating power of the scale.

c. The halo effect. Where several dimensions of the operator's performance are rated separately (e.g., speed, accuracy, coordination), the rating given any single dimension is likely to be affected by the rater's general impression of the individual being rated. Halo effects are most likely to be found in performance attributes that are not clearly (objectively) defined and that are not easily observed.

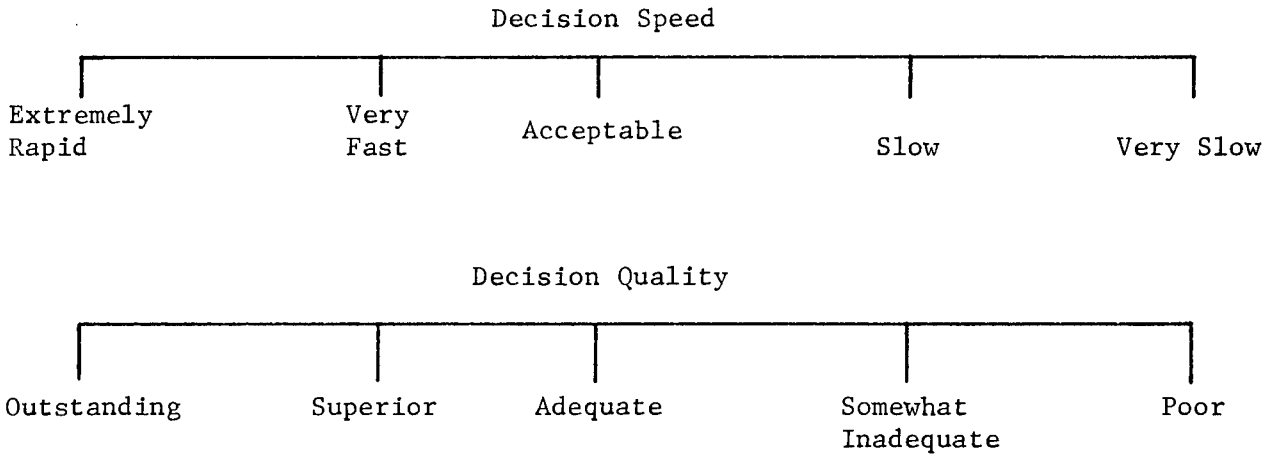
5. Example. Because many of the functions performed in the TACC do not lend themselves to objective evaluation, it was necessary to rely heavily on ratings of performance effectiveness. Theoretically, these would be based on observable criteria for which a judgment could be made. A number of rating scales were developed but only one will be discussed in any detail.

Since the TACC is divided into functional areas that operate semi-independently, an efficiency rating scale had to be developed for each area. In this case, it was the Air Defense Coordinator (ADC) together with his

assistant. The first step in the development of their rating scales was the specification of the functions to be evaluated and the effectiveness criteria for these functions. The ADC functions include monitoring, supervising and coordinating control of aircraft and missiles by subordinate agencies, initiating orders to these agencies, monitoring equipment and operational status, controlling the tactical alert net, etc. Speed and accuracy are obvious effectiveness criteria for these functions, but the problem was that objective values of these criteria did not exist. Attempts to develop standard values for these criteria were ultimately abandoned because agreement on them could not be reached.

It was decided that the one critical element in the ADC's activity that cut across all his functions was his decision making and this involved two aspects: speed and decision quality. This made two scales necessary. Effectiveness was to be determined by the judgment of highly skilled, experienced personnel who had operated an earlier version of the TACC.

The effectiveness rating scales for the ADC as finally developed took the following form:



Because of the lack of objective standards for decision speed and quality, the responsibility for valid measures depended completely on the rater. For this reason, it was decided that any evaluation score less than adequate (midpoint of the scale) would be documented by the rater with specific instances of the behavior on which the evaluation was based. In that sense, the rating was supplemented by critical incident data (see subsequent discussion). These data were evaluated to justify a less than desirable score.

The rater was also required to practice making observations and his first three or four ratings were discarded as essentially practice evaluation. Ideally one would have wished to have two raters simultaneously and independently evaluating ADC performance, but the lack of qualified observers prevented this.

6. Summary. In making use of the rating method the following steps should be implemented:

a. Determine the variables for which ratings will be sought and the continuum on which those variables can be rated.

b. Develop objective criteria to define the range of values along which the variables will be rated.

c. Utilize the rating scale development principles described in this section.

d. Try the rating scale out to see how well it performs.

e. Provide systematic training to the rater in observation of the variables being rated and use of the scale.

The Checklist

1. Definition. The checklist is a series of statements that describe characteristics of (a) operator performance, (b) the equipment configuration, and/or (c) system operations. The checklist statement serves as a standard against which the evaluator makes a judgment that what is being evaluated either has or does not have the characteristic described in the statement.

2. Information Provided. The checklist is used to verify that the system satisfies its requirements. It is used most frequently to evaluate relatively static attributes of the system, such as characteristics of the man-machine interface; one finds it most commonly in the form of the human engineering checklist. The checklist is not often used to describe performance because performance involves continuous variables and the checklist statements are binary. In general, human engineering evaluations of equipment should be performed during developmental testing (prior to OST); however, the occasion may arise to use the checklist method during OST.

3. Use Factors. Like the other methods described previously, the development of a checklist presumes a criterion: that the object or performance being evaluated should possess certain desired characteristics. It is impossible to develop a checklist of qualities before having decided what those qualities should be. The checklist also presumes a value judgment: if the object or performance exhibits characteristics X, Y, and Z, it is adequate or effective in performance; if it does not, it is inadequate and ineffective.

If one were to develop a checklist to evaluate performance, it might contain statements such as the following:

a. Consults TMs frequently during checkout.

b. Strips machine gun quickly and correctly.

These statements suggest that consulting TMs frequently and stripping a machine gun quickly and efficiently are desirable qualities of task performance. What do frequently, quickly, and efficiently mean? Since these adverbs

reflect an underlying quantitative continuum (e.g., TMs might never or always be consulted), they can be measured more directly by means of time and error measures and frequency of occurrence.

On the other hand, if the evaluator uses a checklist because he cannot precisely define what he means by frequently, quickly, and efficiently, then any evaluation he might make using these terms would be highly invalid and inconsistent.

Because dynamic performance can be measured by objective methods, the checklist is not a very efficient means of performance measurement. Where, however, the object of the evaluation is static and where the qualities to be measured are nonquantitative, the checklist can be useful, provided that these qualities as expressed in the checklist are meaningful in terms of some standard or criterion. Moreover, the checklist statement does not indicate how much of that characteristic the object or performance being evaluated possesses. Hence, the checklist is purely a binary device. Because of this quality it may be misleading, since few characteristics exist totally or not at all.

The checklist is most frequently used for evaluating the human engineering characteristics of equipment. Criteria for specifying the desirable characteristics of the man-machine interface can be found in MIL STD 1472B (DoD, 1974)[3] which governs the human engineering design of all military systems. By determining the particular characteristics that represent the man-machine interface under evaluation, one can refer to 1472B and find the features to be described by checklist statements. For example, on page 59 of 1472B one finds the paragraph 5.4.1.2.1, Consistency of Movement, which can be applied to the equipment if it has controls:

> Controls shall be selected so that the direction of movements of the control will be consistent with the related movement of an associated display, equipment component, or vehicle. In general, movement of a control forward, clockwise, to the right, or up or pressing or squeezing a control shall turn the equipment or component on, cause the quantity to increase, or cause the equipment or component to move forward, clockwise, to the right, or up . . . .

The checklist items that can be developed from this paragraph are:

a. The direction of control movements is consistent with the related movement of their associated displays.

b. Movement of a control forward, clockwise, to the right, or up causes the quantity to increase, or causes the equipment/component to move forward, clockwise, to the right, or up.

---

[3]Department of Defense, MIL STD 1472B. Human Engineering Design Criteria for Military Systems, Equipment, and Facilities. 31 December 1974.

4. Problems. Using the checklist requires that its statements be interpreted. Because each statement implies a binary situation (the equipment has or does not have the specified characteristic), intermediate situations present problems. For example, the previous checklist items assume that all controls possess the direction of movement attribute. Suppose, however, that of four controls, three possess the quality and one does not. Should the item be checked off as OK or not OK?

The specificity of the checklist item also determines the precision with which the checklist evaluation can be performed. For example, page 17 of MIL STD 1472B (paragraph 4.6, Simplicity of Design) specifies:

> The equipment shall represent the simplest design consistent with functional requirements and expected service conditions. It shall be capable of operation, maintenance, and repair in its operational environment by personnel with a minimum of training.

Assuming one translates this requirement into a checklist item, it might read as follows: "Equipment design is simple, easy to operate and maintain."

In making this checklist evaluation, the evaluator must ask himself, what does "simple, easy to operate and maintain" mean? Obviously the grossness of these concepts makes it impossible to derive any meaningful results from their evaluation.

The checklist also suffers from another difficulty: its judgments cannot be handled quantitatively. Since the product of the checklist evaluation is merely a "Check" that a specified quality does or does not exist, the most one can do is to sum the number of such checks. Obviously, the greater the number of these, the better (or worse) the system or equipment is; but how much better or worse cannot be determined. The sum of a checklist evaluation cannot, moreover, be added to any other statistic describing performance. It cannot be treated like a rating because it is noncontinuous.

Checklist data are therefore merely descriptive. The checklist does, however, serve to remind the evaluator that certain qualities should be looked for. Once those qualities have been judged, the checklist evaluation pinpoints the characteristics that need improvement.

There is no standard human engineering checklist as such, because the nature of the system attributes to be examined will vary with the system. Because of the large number of qualities which any moderately sized system has, it is possible to develop a highly detailed checklist with hundreds of items; on the other hand, by describing the system more broadly, one can have a very short checklist. The length of the checklist will depend on the number of qualities the evaluator decides are critical to effective system operation; there are no general rules that can be applied.

Many evaluators (in industry) make up their own human engineering checklists, but a number are already available which can be very useful. Notable among these is HEDGE (U.S. Army, 1974),[4] which was specifically developed for the U.S. Army and therefore deals with characteristics of equipment similar to that of the Marine Corps.

5. Example. Human engineering checklists are, of course, applicable to the TACC, principally in terms of the control/display configuration of the situation display consoles, and somewhat less so to the WASDG and the manual status boards. Of these, the most important are the consoles. Among the console characteristics that might be evaluated using the human engineering checklist are the size, legibility, and intensity of symbols on the situation display CRT, the specific symbols and codes provided, and the arrangement of controls for processing information. Adequacy of procedures for performing console operations would also be evaluated using the checklist.

During the developmental testing of the TACC, the individual TACC equipments had been evaluated by means of human engineering checklists (as well as by other means). However, TACC evaluators retained the option of reevaluating these equipments if other evaluation results suggested difficulties in operating the equipment.

6. Summary. In making use of a human engineering checklist the following steps are desirable:

a. Check to determine whether the system being evaluated has previously been analyzed using a human engineering checklist.

b. If a human engineering checklist evaluation is desirable, refer to one of the already developed checklists like HEDGE to see whether it or any of its items adequately describe the system to be evaluated.

c. If already existent checklists are not satisfactory, refer to MIL STD 1472B to determine which sections of this standard apply to the system under evaluation.

d. Select applicable provisions of MIL STD 1472B and translate these statements into checklist form.

e. Prior to performing the human engineering evaluation, select qualified human factors engineers or operational personnel with evaluation experience as the checklist evaluators. Because of the subjectivity involved in such evaluations, it is desirable to have more than one evaluator apply the checklist.

f. Have the checklist evaluator study the system to be evaluated before applying the checklist.

---

[4]U.S. Army. HEDGE (Human Factors Engineering Data Guide for Evaluation). U.S. Army Test and Evaluation Command, Aberdeen, MD, March 1974.

Critical Incidents

1. Definition. A critical incident is an unusual event that occurs during system operation and that illustrates some outstanding (either positive or, more likely, negative) characteristics of the system.

Although it can describe any unprogrammed occurrence (such as an equipment malfunction), the term is generally taken to mean some task-related action performed (or not performed) by the operator. One such instance was recorded by the author many years ago during the testing of an early guided missile which was launched on a test run at a range near the city of Albuquerque, New Mexico. Standing at one of the consoles in the blockhouse over a very prominent red destruct button was the Safety Officer, whose task was to destroy the missile by pressing the button if the missile drifted off-course. Shortly after the flight began, the missile began to drift off-range. The Safety Officer was poised over his console as it became apparent that the vehicle, if it continued its flight, would impact somewhere in Albuquerque. The drift became marked, but the Safety Officer, mesmerized and sweating, did not respond. Just as the missile was about to descend on the city, someone thrust the Safety Officer aside, pressed the destruct button, and destroyed the vehicle.

The example suggests certain characteristics of the critical incident. The action taken (or not taken) is comparatively rare. Because it is unprogrammed, it is difficult to define in advance of its occurrence. It represents either extremely effective or extremely ineffective behavior (more usually the latter, because it is easier to recognize deficiencies), but it is obviously in a class distinct from more characteristic errors. Since it is unprogrammed, it is difficult to set up instrumentation to secure such data; hence, it is generally recorded through human observation, all the more so because it requires interpretation to recognize that the phenomenon is occurring.

2. Information Provided. Since the critical incident is unusual, it may imply an underlying system problem. Since it directly reflects the operator's performance, which includes his training, it may suggest certain changes needed in future training or other personnel requirements.

3. Use Factors. The incident is reported verbatim, almost like a story: who, what, when, where, why (if known), and the consequences (if any) of the incident. Since the event is relatively rare and disproportionate to preceeding events, its interpretation is by no means obvious. Consequently, the cause of the incident should be investigated subsequently by interviewing the performer responsible for the incident. Events occurring previous to the incident can be examined for clues to its meaning.

4. Problem. The critical incident cannot be quantified except by counting the frequency of its occurrence and even this does not tell us much. However, the incident can be highly diagnostic of problem areas deserving further investigation.

5. Example. As was discussed in the section on ratings, the evaluator of the TACC ADC was required to justify any less-than-adequate rating given to the ADC's performance; the justification was a critical incident that demonstrated the inadequacy. As it turned out, only a few such critical incidents were reported (primarily in the early stages of the OST). However, these were important because they occurred as a result of unusual events for which procedures had not yet been developed and hence suggested the development of those procedures.

Summary

1. Whenever possible, use both individual and team measures concurrently in order to determine the contribution of the individual operator to system output.

2. Objective measures are preferable to subjective ones, but both should be used. Subjective methods supply information that helps to explain the meaning of objective data. Any evaluation will be incomplete unless subjective data are gathered in addition to objective data.

3. After selecting a general class of measure (e.g., accuracy), the test planner must develop that measure in terms of the particular task and system he is evaluating.

4. In selecting a measure and measurement method, certain criteria should be applied. The measure should (a) be objective, quantitative, unobtrusive, and easy to collect, (b) require, where possible, no specialized psychological data collection techniques or instrumentation, and (c) cost little or nothing.

5. Reaction time is the time between the occurrence of an event and the start of the action required by the event. If a time requirement does not exist, RT is unlikely to be of value in evaluating the system unless the system is required to respond as quickly as possible.

6. Task duration is important only when a maximum duration is specified for a task or group of tasks. However, this type of measure is easy to collect.

7. All errors are not the same, differing in terms of the criticality of their effects on the task and the system output. Hence, errors should be weighted on a criticality scale. Before collecting error data, an analysis should be performed on the kinds of error likely to occur and their potential effects. A criterion of what constitutes acceptable accuracy should also precede any error measurement. Trivial errors need not be collected.

8. The frequency with which events (e.g., errors) occur should also be recorded because the data may have diagnostic value.

9. The percentage of tasks successfully accomplished by the operator (s/n ratio) is a useful measure of amount achieved by the system, provided the tasks being evaluated are discrete and are performed several times.

10. No performance evaluation should be considered complete without an interview with the performer. Although there are substantial advantages to the interview method (flexibility, interpersonal relationships), it is very costly in terms of subject time.

11. Along with the interview, the questionnaire is the most commonly used subjective method. It covers essentially the same ground as the interview but is less costly, being administered to groups rather than individuals. However, it is less flexible than the interview and more subject to misinterpretation by the respondent.

12. Observation is the means by which all data are collected; however, it is especially important in applying subjective measures, particularly ratings. Observation is extremely difficult to apply correctly unless what is to be observed is defined precisely and training is given the observer.

13. Graphic rating scales are particularly popular in subjective evaluation, because they allow numbers to be assigned to observations. However, rating accuracy depends upon the precision of the observer who is rating, and this in turn depends upon the precision of the rating criterion. The most difficult step in the development of the rating scale is to conceptualize the performance variable being rated in terms of a continuum. Training in using rating scales is essential.

14. Checklists are based on the evaluator's judgment of characteristics of operator performance, equipment or system operation. The judgment is that what is being evaluated either has or does not have the checklist characteristics. The checklist is therefore limited because of its binary nature.

SECTION FOUR--HUMAN ENGINEERING CHECKLIST PROCEDURES

This section describes a human engineering checklist which can be used to evaluate the adequacy of new equipment from the standpoint of operability and maintainability. The checklist in this section was developed to satisfy the following criteria:

1. It should be reasonably short and yet cover the essential features.

2. It should not require either specialist background or instrumentation.

3. Checklist items should be capable of being evaluated by inspecting the physical features of the equipment being evaluated or by simply measuring or observing the equipment while it is being operated.

Introduction

The purpose of this section is to present a human engineering checklist that will enable the evaluator to ensure that equipment features critical to personnel operation and maintenance of equipment have either been included or, in the case of negative features, avoided.

The checklist has been designed for use by operational personnel who are not human engineering specialists. It is based on an earlier checklist developed for the U.S. Army--the Human Factors Engineering Data Guide for Evaluation (HEDGE),[1] which was in turn based on MIL-STD 1472B.[2] The present checklist has been, however, extensively modified for Marine Corps use.

The checklist described in this Section has been based on the following assumptions:

1. If the checklist user is a military man, he should not be asked to employ criteria that demand a specialist background. Consequently, items that require a specialist background have been eliminated.

2. Criteria that are so general that one has difficulty using them for evaluation (e.g., the requirement that the equipment should be designed simply) should be avoided.

3. Relatively few equipment features can be meaningfully observed in making a human engineering evaluation. Long lists of checklist items that follow the provisions of MIL-STD 1472B step-by-step are merely deceptive; even experienced human engineers cannot properly evaluate all these features. Moreover, for many of these checklist items, even when the man-machine system is found to deviate from these criteria, the significance of the deviations (in terms of whether or not they will seriously affect operator/maintainer performance) is suspect.

---

[1]U.S. Army, Human Factors Engineering Data Guide for Evaluation (HEDGE) U.S. Army Test and Evaluation Command, March 1974.

[2]Department of Defense. MIL-STD 1472B, Military Standard, Human Engineering Design Criteria for Military Systems, Equipment and Facilities. Washington, D.C., 31 December 1974.

4. Most of the criteria specified in MIL-STD 1472B (which governs the human engineering requirements of all Department of Defense equipment) are designed as guidance to the engineer and not as evaluation criteria.

The checklist has also been shortened as much as possible to make it more acceptable to users.

Like most checklists, this one is not designed to produce a number that summarizes the "goodness" or "badness" (from a human engineering standpoint) of an equipment. Rather, it is designed to allow the evaluator to pinpoint the equipment features that require changes either in hardware or procedures. In each evaluation, the evaluator must use common sense to decide whether a particular equipment feature deviates sufficiently from the checklist criteria to significantly affect operator/maintainer performance.

In using the checklist, the evaluator must, of course, know how the equipment functions. If he does not, he should start by analyzing the operating (or maintenance) procedure for the equipment and then inspect the equipment to see if it agrees with checklist criteria. For example, suppose the checklist item is "Controls should be located under or to the right of their associated displays." In this case, if the evaluator does not know which controls are associated with which displays, he should go to the operating procedure for the answer. In most cases, if the evaluator is at all familiar with the equipment and its operation, even this preliminary step (analyzing the operating/maintenance procedure) will not be required.

Human Engineering Checklist

Obviously not every checklist item will be pertinent to a particular equipment and the evaluator will have to examine the list of items to see which are relevant to his particular equipment. Checklist criteria in this section are broken out in terms of the following categories:

1. Controls.
2. Displays.
3. Labels.
4. Workspace.
5. Stairs and ladders.
6. Handles, handholds, and railings.
7. Doors, hatches, and entryways.
8. The working environment.
9. Lines and cables.
10. Fasteners and connectors.
11. Cases and covers.
12. Access openings.
13. Maintainability.
14. Communications.

CONTROLS

1. Controls should be located under or to the right of their associated displays.

*2. Control movement should be consistent with the related movement of an associated display, equipment component, or the vehicle as a whole. For example, a control movement forward, clockwise, to the right or up, or depressing the control, should move the equipment forward, etc.

*3. Rotary valve controls should open counterclockwise.

4. Controls should be grouped by common functions or arranged in the operating sequence (if possible).

5. Critical, frequently used controls should be located in the most favorable position for reaching (e.g., lower quarter of panel).

6. Controls used solely for maintenance should be separate from normal operating controls and covered during normal equipment operations.

7. Minimum adjacent edge separation distances for:

   a. Rotary selector switch: 1 hand, 1"; 2 hands, 3".

   b. Thumbwheel, 0.4".

   c. Knob, 1".

   d. Pushbuttons, 0.5".

   e. Toggleswitch, 0.5".

   f. Lever, 1 hand, 2".

   g. Pedals, 4".

   h. Cranks and handwheels, 3".

8. Minimum and maximum sizes of the following controls should be:

CONTROL	HEIGHT	LENGTH	WIDTH	DIAMETER
a. Rotary selector switch	.625-3"	1-4"	1.0"	
b. Knobs	0.5-1"			1-4"
c. Handwheels				2-4.25" (1 hand) 7-21" (2 hands)
d. Pushbuttons				0.385"-0.75"
e. Toggle switches		0.5-2" (Bare) 1.5-2" (Gloved finger)		

Note. Criteria marked with an asterisk can be evaluated by physically manipulating the controls.

9. Shape coding should be used for rotary controls having dissimilar functions. If size coding is used, use no more than 3 sizes.

10. Controls operated sequentially should be arranged left to right and top to bottom.

*11. Control size and separation distance should permit the use of gloves in the operating environment.

12. "Deadman" controls should be used when the operator's incapacity could produce a critical condition.

13. Extremely critical controls should be protected from inadvertent activation by locating them apart from other controls; by recessing, shielding or covering them; and by using interlocks.

14. Control-display relationships should be apparent through proximity of related controls and displays, through similar groupings, by using a common coding scheme, or by framing or labelling related controls/displays.

*15. Neither the control nor the hand used for operating the control should obscure an associated display.

Note. Criteria marked with an asterisk can be evaluated by physically manipulating the controls.

DISPLAYS

1. Only necessary information should be displayed and only to the required degree of detail.

2. Display face should not be less than 45 degrees to operator's line of sight.

3. Functionally related displays should be located close to each other.

4. Critical and frequently read displays should be located in the center of display panel.

*5. Minimum-maximum viewing distances to display panel: 13-28 inches.

6. Absence of signal should not indicate ON, READY, or OK condition.

7. Displays used while performing maintenance should be covered or non-visible during normal equipment operation.

8. Warning lights should be integrated with or adjacent to remedial controls.

Where possible, the warning light should have a flash rate of 3-5 flashes per second, with approximately equal on-off time.

9. Display colors should read:

a. Steady red for failure or NO GO.

b. Flashing red: emergency.

c. Yellow: marginal.

d. Green: OK.

Avoid blue.

10. Scale markings should be linear and should progress by 1, 2, or 5 units; intermediate marks should not exceed 9.

When positive and negative values are displayed around a zero position, the zero point should be at 12 or 9 o'clock.

11. Pointers should be close to dials to avoid parallax; they should not obscure or exceed width of index marks.

Note. Criteria marked with an asterisk can be evaluated by physically moving the observer. Criterion Number 1 can be evaluated by interviewing operators.

12. Master caution, warning, or summation lights should be set off by themselves.

13. If the face of the scale is color coded, green represents the operating range; yellow, caution; and red, danger.

14. Scales should read clockwise, left to right, or from bottom up; numbers should be oriented upright and placed outside markings.

*15. CRT displays: viewing distance should be approximately 16 inches; they should be hooded or shielded when detection of faint signals is required.

16. Audio displays should be used only when operator vision is overburdened, speech is required, the operator is occupied elsewhere or cannot move his body, or where redundancy or warning signals are required.

17. To check on audio display parameters, see a Human Factors specialist.

18. Audio warning signals: should automatically reset; minimum duration 0.5 second.

19. To check on microphone and headphone parameters, see Human Factors specialist.

20. In high noise environment, use binaural headsets capable of reducing ambient noise to comfortable level.

Note. Criteria marked with an asterisk can be evaluated by physically moving the observer.

LABELS

1. Labels should be placed above controls and displays.

2. Labels should ordinarily be oriented horizontally; vertically only when not critical for personnel safety or performance.

3. Labels should be placed on or near items they identify; ensure that they do not obscure other information and are not obscured by other equipment components.

4. Labels should describe the functions of the units they identify.

5. Labels should be written with standard abbreviations and familiar words only; ensure that trade names are not used.

6. Character height and viewing distance should be in accordance with the following:

VIEWING DISTANCE	HEIGHT
20" or less	0.09"
20-36"	0.17"
36-72"	0.34"
72-144"	0.68"
144-240"	1.13"

7. Conspicuous placards should be placed adjacent to hazardous equipment.

8. Pipe, hose, and tube lines should be clearly labelled as to contents and pressure temperature; electrical receptacles should be marked with voltage, phase, and frequency.

WORKSPACE

1. Displays to be read when standing should be 41-74" above the floor; to be read when seated, between 6-48" above the seat.

2. Controls to be operated when standing should be 34-74" above the floor; to be operated when seated, 8-35" above the seat.

3. Seats should be vertically adjustable from 16-21" in 1" increments. The supporting backrest should be reclinable 103-115 degrees. Arm rests should be provided where feasible.

4. The operator should have at least 4 feet of free space in front of his console.

5. The operator should have the following minimum knee room at his work station: 25" high, 20" wide, 18" deep.

6. Work surface width should be at least 30" wide and 16" deep.

7. Work station table tops should be 29-31" above the floor.

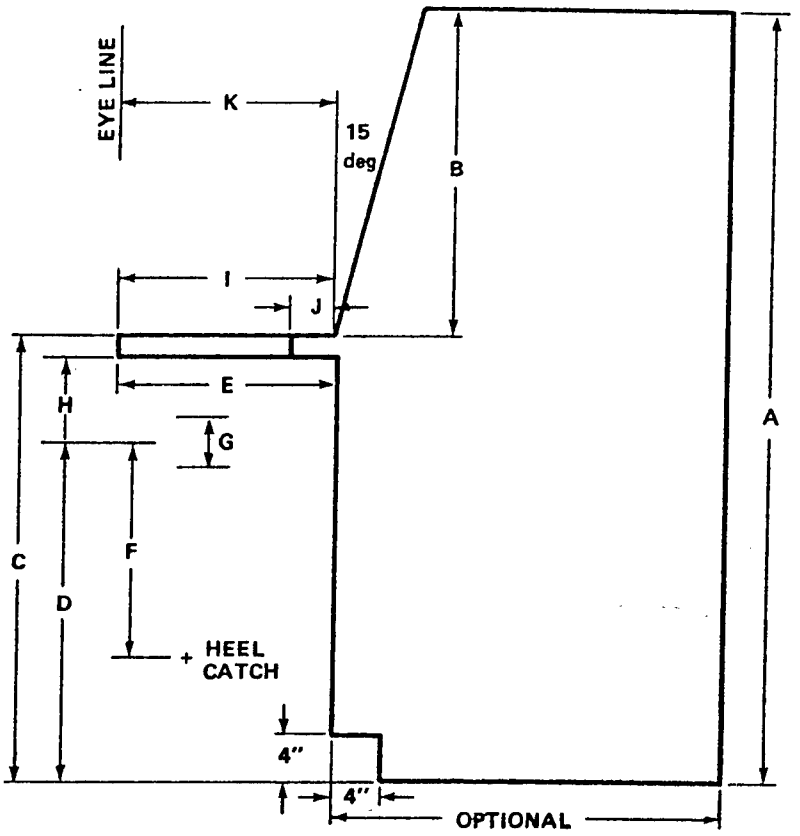
8. Console design should accord with Figure 4-1.

9. Seat design should accord with Figure 4-2.

10. Seats should be large enough to accommodate suitably clothed personnel.

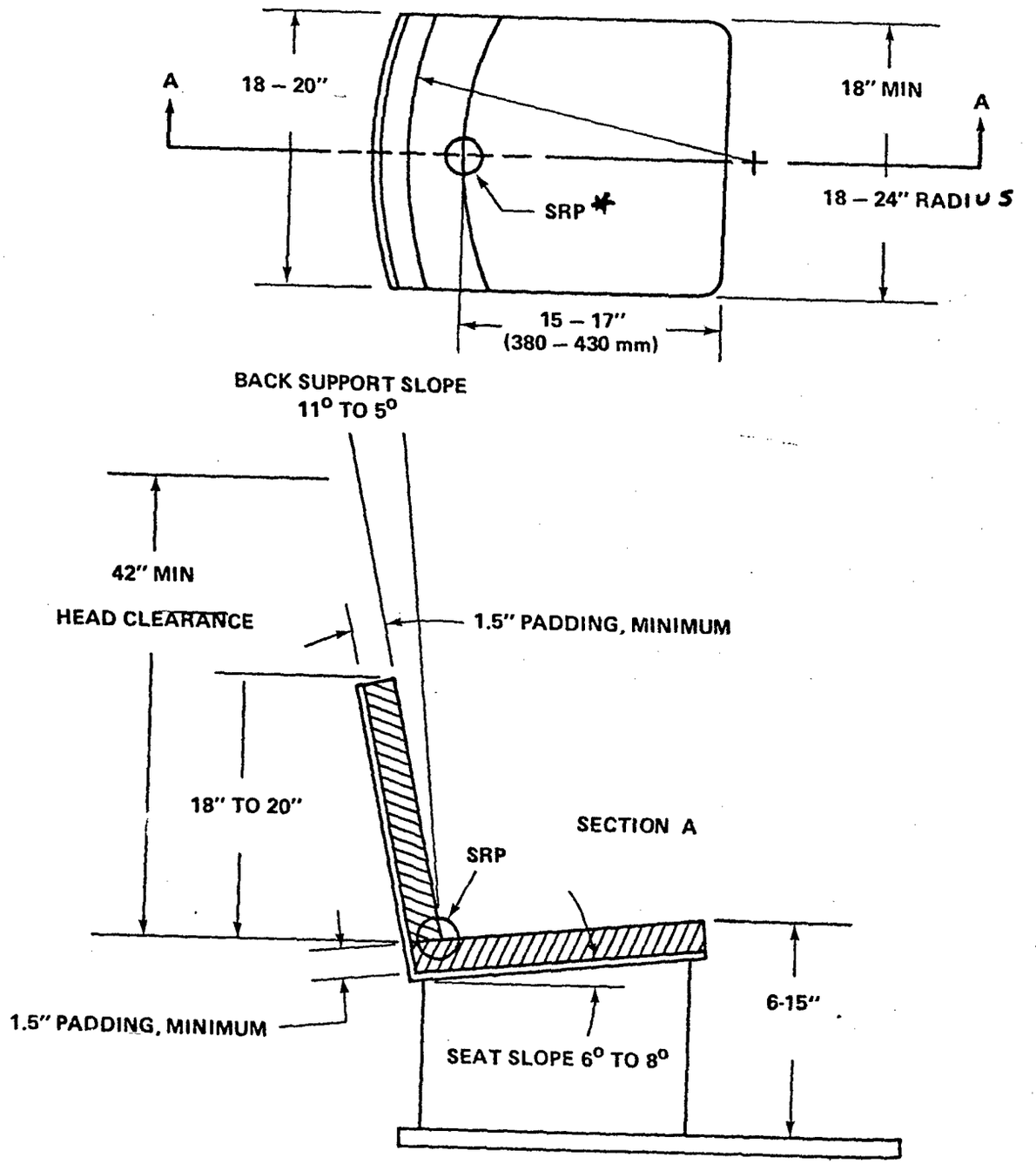
11. The work station should have adequate storage space for paperwork, spare parts, etc.

12. Emergency escape facilities should be provided and marked.



KEY	DIMENSIONS	INCHES
A	MAXIMUM TOTAL CONSOLE HEIGHT FROM STANDING SURFACE	51.5 - 58"
B	SUGGESTED VERTICAL DIMENSION OF PANEL, INCL SILLS	26"
C	WRITING SURFACE: SHELF HEIGHT FROM FLOOR	25.5 - 36"
D	SEAT HEIGHT FROM STANDING SURFACE AT MIDPOINT OF "G"	18 - 28.5"
E	MINIMUM KNEE CLEARANCE	18"
F	FOOT SUPPORT TO SITTING SURFACE	18"
G	SEAT ADJUSTABILITY	5"
H	MINIMUM THIGH CLEARANCE AT MIDPOINT OF "G"	6.5"
I	WRITING SURFACE DEPTH INCLUDING SHELF	16"
J	MINIMUM SHELF DEPTH	4"
K	EYE LINE-TO-CONSOLE FRONT DISTANCE	16"

Figure 4-1. Standard console dimensions key.



* SRP = SEAT REFERENCE POINT

Figure 4-2. Dimensions for vehicle operator's seat.

STAIRS AND LADDERS

1. Fixed ladders should have the following dimensions:

   a. Rung thickness: .75-1.50".

   b. Rung spacing: 9-16".

   c. Ladder width: 12" minimum, 18-21" recommended.

   d. Ladder depth: 30-36".

   e. Space behind ladder: 6" minimum.

   f. Height of ladder above landing: 33" minimun.

2. Toe board or guard screen around platforms should be 3" minimum height.

3. Guard rails should be provided around the opening at the tops of fixed ladders.

4. Adequate footholds should be provided for crew to reach vehicle hatch from ground.

5. Appropriate labels warning of possible hazards should be posted.

6. The exterior ladder tread of personnel platforms should be open grating or nonskid treated.

HANDLES, HANDHOLDS, AND RAILINGS

1. Handles or grasp areas should be located at the equipment's center of gravity and at least 2" from any obstructions.

2. Handles should be recessed where practicable.

*3. Fingers should be able to curve around the handle comfortably.

4. Handles should have minimum 2" clearance during use.

5. Stairs, ladders, platforms, and ramps should be equipped with a handrail on each side.

*6. Handholds should be furnished where needed and within easy reach; they should accommodate gloved hands.

7. The handrail diameter should be 1.25-3".

8. Guardrails for personnel platforms should have a top rail height of not less than 42". The distance between the platform edge and the centerline of the railing should not exceed 2.5".

9. Handrail to handrail width on stair ladders should be 21-24"; on one-way stairs, 30" minimum; on two-way stairs, 48" minimum.

10. Handrail clearance from the wall should be a minimum of 1.75" for stairs; for stair ladders, 2".

11. Handrails should have nonslip surfaces.

Note. Criteria marked with an asterisk can be evaluated by physically performing required actions.

DOORS, HATCHES, AND ENTRYWAYS

1. Wall hatches should be flush with the floor where structurally possible.

2. Round hatches should have a minimum 30" diameter.

3. Rectangular hatches to be used by personnel with light clothing should be (for top and bottom access) 23" wide, 13" high; for side access, 30" wide, 26" high. If used by personnel with bulky clothing, they should be (for top and bottom access) 27" wide, 16" high; for side access they should be 34" wide, 29" high.

4. Floor escape hatches should have a minimum 22" diameter.

5. Latch handles should release down or forward; catch and lock, up or aft. They should be operable with gloved hands.

*6. Emergency doors should be quick opening.

*7. Escape openings should be free of obstructions and permit passage of personnel with survival gear.

Note. Criteria marked with an asterisk can be evaluated by physically performing required actions.

ENVIRONMENT

1. Temperature in a moving vehicle should be 50 degrees (F) minimum, 85 degrees (F) maximum. Above the maximum temperature, air conditioning should be provided.

2. Noise level should permit necessary person-to-person and telephone communication without shouting. If specific measures of noise level are required, see paragraph 5.8.3 and following of MIL-STD 1472B and refer to a Human Engineering specialist.

3. If specific measures of humidity are required, see paragraph 5.8.1.4 of MIL-STD 1472B and refer to Human Engineering specialist.

4. If specific measures of illumination are required, see Table XIX, pp. 141-144 of MIL-STD 1472B, and refer to Human Engineering specialist.

5. If specific measures of vibration are required, see Figure 36, pp. 149 of MIL-STD 1472B, and refer to Human Engineering specialist.

6. The lateral field of view in a moving vehicle should be 180 degrees minimum. The vehicle operator should be able to view the ground at all distances beyond 10 feet in front of the vehicle. Upward visibility should be 15 degrees above the horizontal.

Note. The adequacy of the work environment can also be ascertained by interviewing personnel and/or by asking them to rate the adequacy of the environment. For rating scale development, see Section Three of this report.

LINES AND CABLES

1. Cables should be routed so as to be accessible for maintenance.

2. Test cables terminating on control panels should not interfere with controls and displays.

3. Cables routed through holes in metal should be protected from mechanical damage by grommets or other protective devices.

4. Cables used for checking units should be long enough for the purpose.

5. When cable clamps are used, they should be spaced approximately every 12".

6. Gas, fluid, and electrical conduit lines should be properly identified.

7. Cables should be labeled to indicate the equipment they are used with and the connectors with which they are mated.

8. Cables should be routed so as not to be pinched by doors, walked on, used for hand holds, or bent.

9. Cables containing individual insulated conductors in a common sheath should be coded.

FASTENERS AND CONNECTORS

1. There should be a 1" minimum space between connectors for grasping.

2. Connecting plugs and receptacles should be color-coded.

*3. Plugs of one voltage should be incapable of being inserted into receptacles of another voltage.

4. Noninterchangeable connectors should be used for different uses.

*5. Fasteners should require only one complete clockwise turn to tighten and one complete counterclockwise turn to loosen.

6. Use identical screw/bolt heads where possible.

7. Aligning pins should extend beyond electrical pins to ensure alignment before pins engage.

8. Use stripes, arrows, etc. to show position of aligning pins.

Note. Criteria marked with an asterisk can be evaluated by physically performing required actions.

CASES AND COVERS

1. Cases should be sufficiently larger than the units they cover to prevent damage when the case is removed and replaced.

2. It should be obvious when a cover is not secured even though it is in place.

*3. Cases should be capable of being lifted from units rather than units lifted from cases.

4. Edges and corners on cases and covers should be rounded or otherwise finished to prevent injury to personnel.

5. Guides, tracks, and stops should be provided as necessary to facilitate handling and prevent damage or injury.

6. If the method of opening a cover is not obvious, instructions should be prominently displayed on cover.

7. Proper orientation of a unit in a case should be obvious through design or labels.

8. Stowage locations should be labeled.

9. Bulkhead, brackets, and other units should not interfere with removal or opening of covers.

10. Mounting screw holes in covers that attach to the chassis should be large enough so that perfect case alignment is not necessary.

Note. Criteria marked with an asterisk can be evaluated by physically performing required actions.

ACCESS OPENINGS

1. When possible, an access should be available whenever frequent maintenance operations would otherwise require removing a case, opening a fitting, or dismantling a component.

2. Size and shape of openings for physical access should agree approximately with dimensions in Figure 4-3.

3. Access covers that are not completely removable should be self-supporting when opened.

4. Accesses and covers should avoid sharp edges to preclude injury.

5. Accesses should be labeled to indicate items to be accessed, operations to be accomplished, and any hazards beyond access.

6. Access warnings should be large enough to be read at a reasonable distance.

*7. Space for gloved hand or clothed body should be provided in access.

8. Access covers should be equipped with grasp areas for openings.

*9. Accesses should be large enough to permit required operations.

Note. Criteria marked with an asterisk can be evaluated by physically performing required actions.

MINIMAL TWO-HAND ACCESS OPENINGS WITHOUT VISUAL ACCESS

Reaching with both hands to depth of 6 to 19.25 inches:

Light clothing:	Width:	8" or the depth of reach*
	Height:	5"
Arctic clothing:	Width:	6" plus 3 4 the depth of reach
	Height:	7"

Reaching full arm's length (to shoulders) with both arms:

Width:	19.5"
Height:	5"

Inserting box grasped by handles on the front:

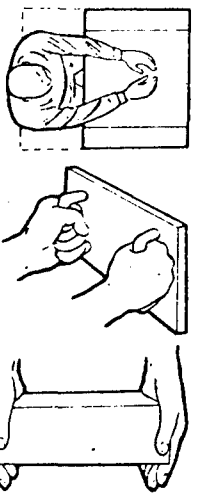
1.2" clearance around box, assuming adequate clearance around handles

Inserting box with hands on the sides:

Light clothing:	Width:	Box plus 4.5"
	‡Height:	5" or 0.5" around box*
Arctic clothing:	Width:	Box plus 7"
	‡Height:	8.5" or 0.5" around box*

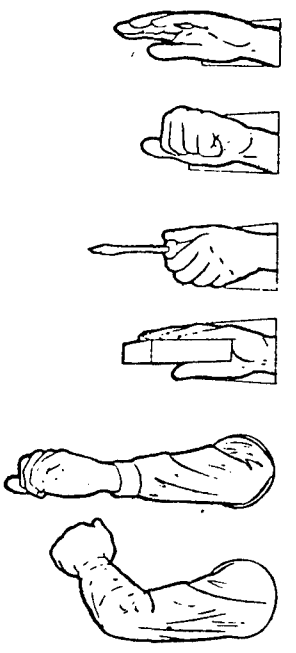
* Whichever is larger.

‡ If hands curl around bottom, allow an extra 1.5" for light clothing, 3" for arctic clothing.



MINIMAL ONE-HAND ACCESS OPENINGS WITHOUT VISUAL ACCESS

	Height	Width
Empty hand, to wrist:		
Bare hand, rolled:	3.75" sq or dia	
Bare hand, flat:	2.25" x	4.0" or 4.0" dia
Glove or mitten:	4.0" x	6.0" or 6.0" dia
Arctic mitten:	5.0" x	6.5" or 6.5" dia
Clenched hand, to wrist:		
Bare hand:	3.5" x	5.0" or 5.0" dia
Glove or mitten:	4.5" x	6.0" or 6.0" dia
Arctic mitten:	7.0" x	8.5" or 8.5" dia
Hand plus 1" dia object, to wrist:		
Bare hand:	3.75" sq or dia	
Gloved hand:	6.0" sq or dia	
Arctic mitten:	7.0" sq or dia	
Hand plus object over 1" in dia, to wrist:		
Bare hand:	1.75" clearance around object	
Glove or mitten:	2.5" clearance around object	
Arctic mitten:	3.5" clearance around object	
Arm to elbow:		
Light clothing:	4.0" x	4.5" or 4.5" dia
Arctic clothing:	7.0" sq or dia	
With object:	Clearances as above	
Arm to shoulder:		
Light clothing:	5.0" sq or dia	
Arctic clothing:	8.5" sq or dia	
With object:	Clearances as above	



MINIMAL FINGER ACCESS TO FIRST JOINT

Push button access:	Bare hand:	1.25" dia
	Gloved hand:	1.5" dia
Two finger twist access:	Bare hand:	object plus 2.0" dia
	Gloved hand:	object plus 2.5" dia

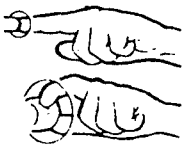


Figure 4-3. Arm and hand access dimensions.

MAINTAINABILITY (TOOLS, TEST EQUIPMENT, TEST POINTS)

1. Special tools should be securely mounted in equipment and accessible to the technician.

2. Test points should be accessible, clearly marked, and close to the units with which they are used.

3. Space should be provided within portable test equipment to store leads, probes, manuals, and tools.

4. Calibration and adjustment controls with limited motion should have mechanical stops to prevent damage.

5. Test points, cables, and connectors should be accessible and visible during maintenance.

6. If nonvisual screwdriver adjustments are required, they should have shaft guides.

7. Displays to indicate failure of equipment units should be provided.

*8. Lamps and light bulbs should be removable from front of display panel without special tools.

9. Sensitive adjustment points should be guarded against accidental disturbance.

10. Larger units should not be mounted to deny access to small ones.

11. Positive and negative battery terminals should be of different size and marked "+" and "-."

12. Items frequently removed for test should, where possible, be mounted on rollout racks, slides, or hinges.

13. Lamp replacement should be possible with power on and without danger.

14. Critical units requiring fast maintenance should be more accessible than other units except that, where criticality is not a factor, units requiring most frequent access should be most accessible.

15. Field removable units should be replaceable with common hand tools.

16. Where applicable, interlocks should be provided to disconnect equipment that would otherwise be damaged by withdrawal of racks or drawers.

Notes. Maintainability deficiencies can also be ascertained by asking personnel to perform required maintenance actions and then interviewing them and/or by asking them to rate the adequacy of maintainability features. For rating scale development, see Section Three of this report.

Criteria marked with an asterisk can be evaluated by physically performing required actions.

COMMUNICATIONS

1. Communication devices should be located within easy reach of the work station.

2. Foot control of the communications device should be available for the seated operator who needs both hands.

3. Headsets should be provided for high noise workspaces.

*4. The speaker should hear his own voice in the headset in phase with his speech.

5. Noise cancelling or bone conduction microphones should be utilized in high noise environments.

6. Microphones, headphones, and telphone headsets should permit normal hands-free operation.

*7. If, in actual use, the operator finds that the volume permitted by his device is too low for him to communicate easily or if he notes any distortion of his speech or of the message received, he should communicate with a Human Engineering specialist.

Notes: Communications adequacy can also be ascertained by interviewing personnel and/or by asking them to rate the adequacy of communications practice. For rating scale development, see Section Two of this report.

Criteria marked with an asterisk can be evaluated by physically performing required actions.

SECTION FIVE--SELF REPORT RATING SCALES

This section describes rating scales which can be used to elicit quantitative judgments from test personnel concerning the adequacy of various aspects of the equipment/system being tested.

Introduction

Among the techniques available to the Marine Corps evaluator for gathering personnel performance data are those involving subjective personnel reactions to characteristics of the equipment/system being tested.

Such subjective data can supplement and amplify objective data. Where objective data cannot be secured, as, for example, to determine an operator's reaction to vehicle driving or riding qualities, subjective methods may be the only evaluation methods available.

This section describes a number of ways of gathering subjective data in quantitative form, principally by means of rating scales to be completed by test participants at the conclusion of a test operation.

Rating Scales

The scales described report the operator's reactions to the following system characteristics:

1. Environmental conditions (noise, temperature/humidity, vibration).
2. Illumination.
3. Handling (driving) qualities of a vehicle.
4. Riding qualities of a vehicle.
5. Control accessibility.
6. Display readability.
7. Control-display arrangement.
8. Information presented.
9. Workspace.
10. External visibility.
11. Vehicle entrance/exit.
12. Accessibility of internal components.
13. Ease of troubleshooting malfunctioning equipment.

14. Test points.

15. General equipment maintainability.

16. Safety.

17. Operating procedures and/or technical manuals.

18. Workload.

19. Communications.

In addition, a Critical Incident Report form and a Satisfaction Checklist are described.

The scales take advantage of the equipment operator/maintainer's experience with the system under test and his general background on comparable systems to tap his evaluation of the system and of personnel performance in utilizing the system. Any or all of these scales can be included in any questionnaire administered to OST personnel or they can be utilized as part of an interview with these personnel. They require minimal explanation to respondents and can be completed very quickly. The evaluator can select his scales to investigate those system aspects about which he wishes information. It is not necessary that all the scales described in this report be used in the same test evaluation.

The evaluator is interested primarily in having the forms available, but he may also wish to know the theoretical foundation of these scales. Information on this point is provided in a Rationale which follows each scale. In a number of scales, the underlying dimension is the amount of effort the operator has to expend in doing his job as related to the particular equipment/system characteristic under investigation. In others, the scale dimension may be the difficulty the operator has experienced in relation to the factor (e.g., workspace) being evaluated. A few scales have more than one dimension, each one contributing to the scale value.

The scales are designed to supply a number representing the operator's evaluation. They therefore differ from the checklist evaluation performed by someone other than the test participant. These scales can be used to supplement the Human Engineering Checklist described in Section Four of this report which was designed to be used by the evaluator himself.

The scales in this report are oriented vertically, rather than horizontally across the paper. This is because a horizontal orientation (which saves paper) tends to crowd the written ratings and descriptors unduly. If space in the questionnaire is a desideratum, the scales can be reoriented horizontally.

Since few operators can differentiate more than five major intervals on a continuum, the scales have five points representing a continuum ranging from Excellent (1.0) through Good (2.5), Fair (4.0), and Poor (5.5) to Unacceptable (7.0). Each point is identified by a behavioral descriptor which "anchors" the point. Intermediate intervals between anchor points (1.5, 2.0, 3.0, 3.5, 4.5, 5.0, 6.0, 6.5) are indicated without being numbered. Test personnel are asked to check anywhere along the scale, paying particular attention to the behavioral descriptors; the resultant checks can easily be transformed into numerical equivalents. Values between anchor and intermediate points are interpolated visually.

The advantage of having a numerical rating of the operator's response to various system characteristics is that the evaluator can treat these ratings statistically (averaging several operator ratings, determining their variability by means of a standard deviation, and comparing the mean ratings of one equipment or one test condition with another).

1. REACTIONS TO ENVIRONMENTAL CONDITIONS (Noise, temperature/humidity, vibration)

RATING ITEM

(Noise, temperature/humidity, vibration)* affected my performance during my operation of the vehicle/equipment, such that

Rating		Descriptor
Not at all	1.0	No discomfort noted; no increase in effort required; no performance impairment.
Slightly	2.5	Minimal discomfort; slight increased effort to perform tasks; minimal performance impairment.
Moderately	4.0	Moderate increase in effort required and/or some discomfort noted; some performance impairment.
Seriously	5.5	Considerable increase in effort required to perform tasks; great discomfort; considerable performance impairment.
Excessively	7.0	Maximum effort required to perform tasks and/or extraordinary discomfort; serious performance impairment.

If rating is seriously or excessively, please comment further.

---

*Select one as appropriate.

Rationale: This item would be used to determine the impact of any undesirable environmental condition on the vehicle/equipment operator. Instrumentation can determine whether any of these conditions would be painful or even damaging to the operator; however, the effect on the operator's performance (short of these extreme conditions) can be most easily determined by the operator's self report. This scale has three dimensions: the amount of (1) effort required to perform the tasks, (2) discomfort experienced, and (3) performance impairment noted. It is assumed that undesirable environmental conditions will increase all three dimensions.

2. ILLUMINATION

RATING ITEM

Because of the illumination within the vehicle or operating compartment, tasks requiring fine visual discrimination

Rating		Descriptor
Excellent	1.0	Can be performed effortlessly.
Good	2.5	Can be performed with only slight effort.
Fair	4.0	Can be performed with moderate effort.
Poor	5.5	Can be performed only with considerable effort.
Unacceptable	7.0	Almost impossible to perform because of inadequate lighting.

If rating is poor or unacceptable, please comment further.

Rationale: This item has essentially the same rationale as item 1. The absolute level of illumination in the vehicle or operating compartment can be measured, but the effect of that illumination on the operator himself can be most readily determined by his self report. (One could of course set up an experiment to measure his performance under different levels of illumination, but this is usually not possible under OST conditions.) The emphasis in this scale is on tasks requiring fine visual discrimination, such as reading dials. The dimension employed in this scale is the amount of extra effort required of the operator by lack of proper illumination.

3. HANDLING (DRIVING) QUALITIES OF VEHICLE

RATING ITEM

The handling (driving) qualities* of my (aircraft, jeep, APC, etc.) ** were such that

Rating		Descriptor
Excellent	1.0	Little or no effort is required.
Good	2.5	Slight amount of effort is required.
Fair	4.0	Moderate amount of effort is required.
Poor	5.5	Considerable effort is required.
Unacceptable	7.0	Very strenuous effort is required.

If rating is poor or unacceptable, please comment further.

---

*Defined as ease of turning vehicle, starting/stopping, shifting gears, etc.
**Insert appropriate term.

Rationale: The dimension in the scale is again the amount of effort or additional effort required to drive the vehicle because of handling deficiencies. It is assumed that, when more than moderate effort is required, the vehicle design is poor or unacceptable. The less effort required, the better designed the vehicle.

4. RIDING QUALITIES OF VEHICLE

RATING ITEM

The riding qualities* of my (aircraft, jeep, APC, etc.)** were such that the ride was

Rating		Descriptor
Excellent	1.0	Always very comfortable.
Good	2.5	Generally comfortable; occasional slight bumpiness.
Fair	4.0	Occasionally very bumpy.
Poor	5.5	Always very bumpy; seat belts required.
Unacceptable	7.0	Heavy pitching/rolling; almost impossible to remain seated during ride despite seat belts.

If rating is poor or unacceptable, please comment further.

---

*Defined as smoothness of ride, sway, vibration, rattles, etc.
**Insert appropriate term.

Rationale: Riding must be differentiated from handling the vehicle. This scale describes what the driver or the passenger feels as a result of vehicle motion apart from efforts made to control the vehicle (item 3). Comfort (and its reverse, bumpiness) is the scale dimension and is assumed to affect operational performance. Since comfort is a subjective response, it cannot be measured objectively.

5. CONTROL ACCESSIBILITY

RATING ITEM

Were controls reachable when you were normally seated?

Rating		Descriptor
Excellent	1.0	All controls are reachable without effort.
Good	2.5	A few controls require slight effort to reach.
Fair	4.0	All controls are reachable but with some additional effort.
Poor	5.5	Considerable straining required to reach a few controls.
Unacceptable	7.0	All controls require considerable straining to reach.

If rating is poor or unacceptable, please comment further.

Rationale: This scale is based on two dimensions: the additional effort required to reach for controls, and the number of controls for which the effort is demanded. The more effort required to reach more controls, the less acceptable the control accessibility. It is possible to determine accessibility objectively by seating the operator, asking him to touch each control in turn, and noting the degree of muscular strain. However, this situation does not impose the same demand on the operator as does normal operations. The scale is designed to measure accessibility in the latter situation.

6. DISPLAY READABILITY

RATING ITEM

Were all displays readable from the normal operating position?

Rating		Descriptor
Excellent	1.0	All displays readable without effort.
Good	2.5	One or two displays require slight additional effort to read.
Fair	4.0	All displays readable but some eye straining required.
Poor	5.5	Intensive eye straining required to read some displays.
Unacceptable	7.0	All displays difficult to read even with intensive straining.

If rating is poor or unacceptable, please comment further.

Rationale: Readability is the ability to discriminate individual characters/numerals on the display. This scale includes two dimensions: amount of effort and number of displays. The greater the amount of eye strain and the more displays for which this eye strain is required, the less acceptable the display readability. Again, this factor can be measured objectively, but only with great difficulty in an OST context.

7. CONTROL-DISPLAY ARRANGEMENT

RATING ITEM

The way controls and displays are arranged on the equipment console is such that operating them was

Rating		Descriptor
Excellent	1.0	Extremely easy and effortless.
Good	2.5	Easy but required a little effort.
Fair	4.0	Slightly difficult, required moderate effort.
Poor	5.5	Very difficult, required considerable effort.
Unacceptable	7.0	Excessively difficult, required strenuous effort.

If rating is poor or unacceptable, please comment further.

Rationale: This scale assumes that the manner in which controls and displays are arranged affects equipment operation, in particular the ease or difficulty of that operation. To use the scale, the respondent must consider control-display arrangement in terms of two dimensions: the amount of effort required and the difficulty he experiences in manipulating those controls/displays. Objective determination of control-display arrangement is very difficult in an OST context.

8. INFORMATION PRESENTED

RATING ITEM

Information presented by displays was

Rating		Descriptor
Excellent	1.0	Extremely easy to understand, required almost no effort.
Good	2.5	Easy to understand but required a little effort.
Fair	4.0	Somewhat difficult to understand, required moderate effort.
Poor	5.5	Very difficult to understand, required very great effort.
Unacceptable	7.0	Either extremely difficult to understand or not enough information is available.

If rating is poor or unacceptable, please comment further.

Rationale: This scale contains two dimensions: ease of understanding the material communicated and the effort involved in doing so. The easier the understanding, the less effort, the better. It is assumed that all necessary information is being presented; when this is not true, the rating becomes Unacceptable.

The following checklist may be used with the preceding scale, or separately.

Check one or more of the following if they pertain to the information displayed on your equipment.

_____ Too much information presented at one time.

_____ Too much information must be combined from different displays.

_____ Information appears too quickly.

_____ Information changes too quickly.

_____ Some information is irrelevant to task.

_____ Not enough information.

9. WORKSPACE

RATING ITEM

Workspace within the vehicle or ground facility was such that there was

Rating		Descriptor
Excellent	1.0	No interference with others and no difficulty in performing own tasks because of space constraints.
Good	2.5	Very infrequent interference with others and/or slight difficulty in performing own tasks because of space constraints.
Fair	4.0	Occasional interference with others and/or moderate difficulty in performing own tasks because of space constraints.
Poor	5.5	Repeated interference with others and/or great difficulty in performing own tasks because of space constraints.
Unacceptable	7.0	Constant interference with others and excessive difficulty in performing own tasks because of space constraints.

If rating is poor or unacceptable, please comment further.

Rationale: Workspace is space available for performing jobs. It is assumed that, if workspace is restricted, the operator will interfere with or be interfered with by others in the same vehicle or facility and he will have difficulty in performing his tasks. Thus, there are two dimensions in this scale: frequency of interference and task performance difficulty. Again, this factor can be measured objectively, but only with great difficulty in the OST context.

10. EXTERNAL VISIBILITY

RATING ITEM

Visibility external to the vehicle was such that I could see out in every required direction

Rating		Descriptor
Excellent	1.0	Without effort.
Good	2.5	With only slight effort.
Fair	4.0	With moderate effort.
Poor	5.5	With great effort.
Unacceptable	7.0	With exceptional effort.

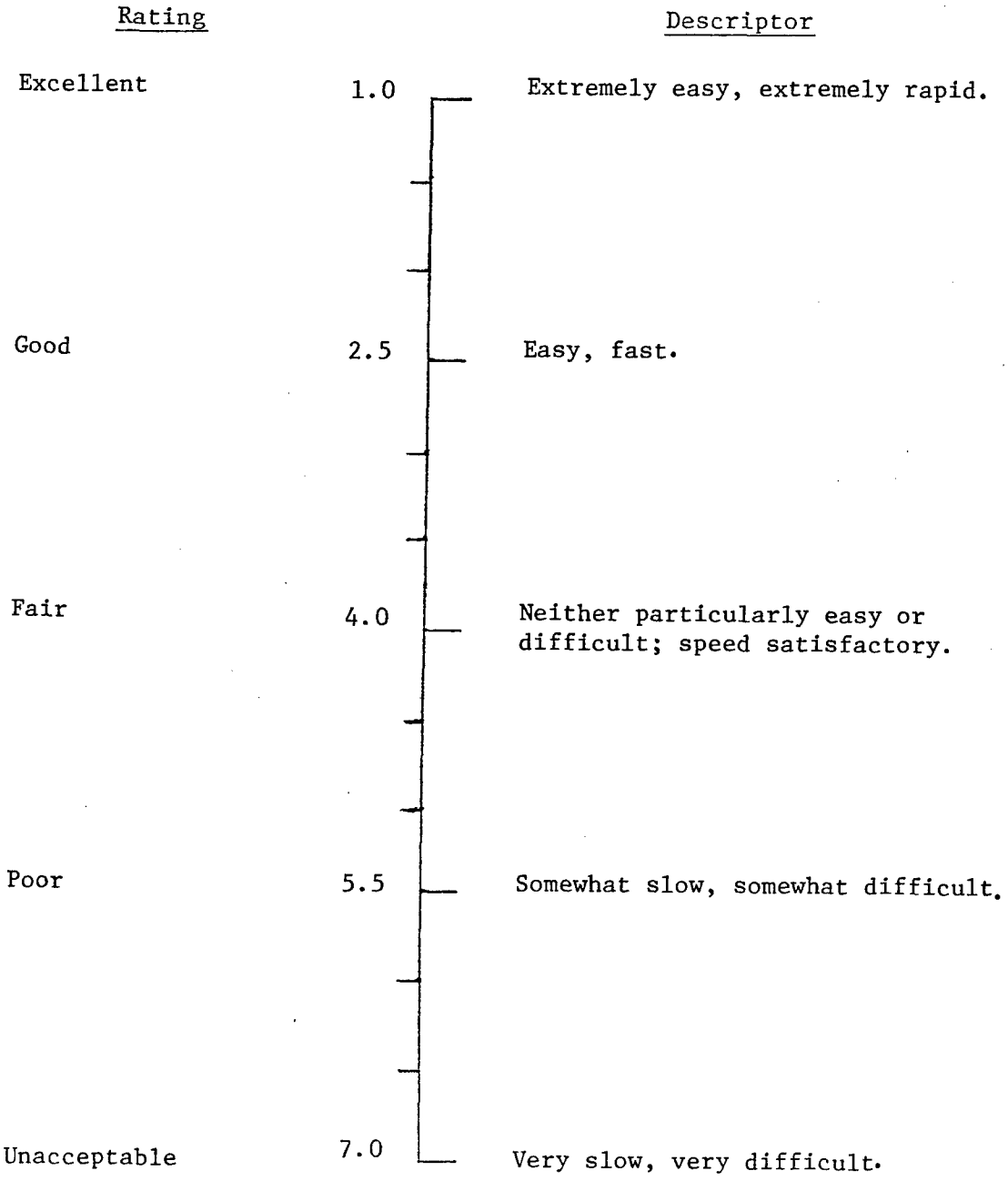
If rating is poor or unacceptable, please comment further.

Rationale: External visibility is defined as how much the operator can see out of windows or viewing ports. It is assumed that, if external visibility is limited but the operator must see outside to do his job, the more restricted the visibility, the more effort he will have to expend on viewing. Again, this factor can be measured objectively, but with great difficulty.

11. VEHICLE ENTRANCE/EXIT

RATING ITEM

Entrance to/exit from the vehicle in full combat gear is



If rating is poor or unacceptable, please comment further.

Rationale: From a performance standpoint, adequacy of entrance to and exit from a vehicle is determined by the speed and difficulty of performing this function. Hence, these two dimensions are included in this scale. It is possible to observe personnel entering/exiting the vehicle and to measure the time required to perform this function. The scale above provides an alternative to this procedure.

12. ACCESSIBILITY OF INTERNAL COMPONENTS

RATING ITEM

Internal components can be reached

Rating		Descriptor
Excellent	1.0	Without effort and without having to remove other components first.
Good	2.5	With only slight effort and after removing only a few other components first.
Fair	4.0	With moderate effort but minimal difficulty; a moderate number of other components must be removed first.
Poor	5.5	With some difficulty; many components must be removed first.
Unacceptable	7.0	Only with great effort/difficulty and after removing an excessive number of other components first.

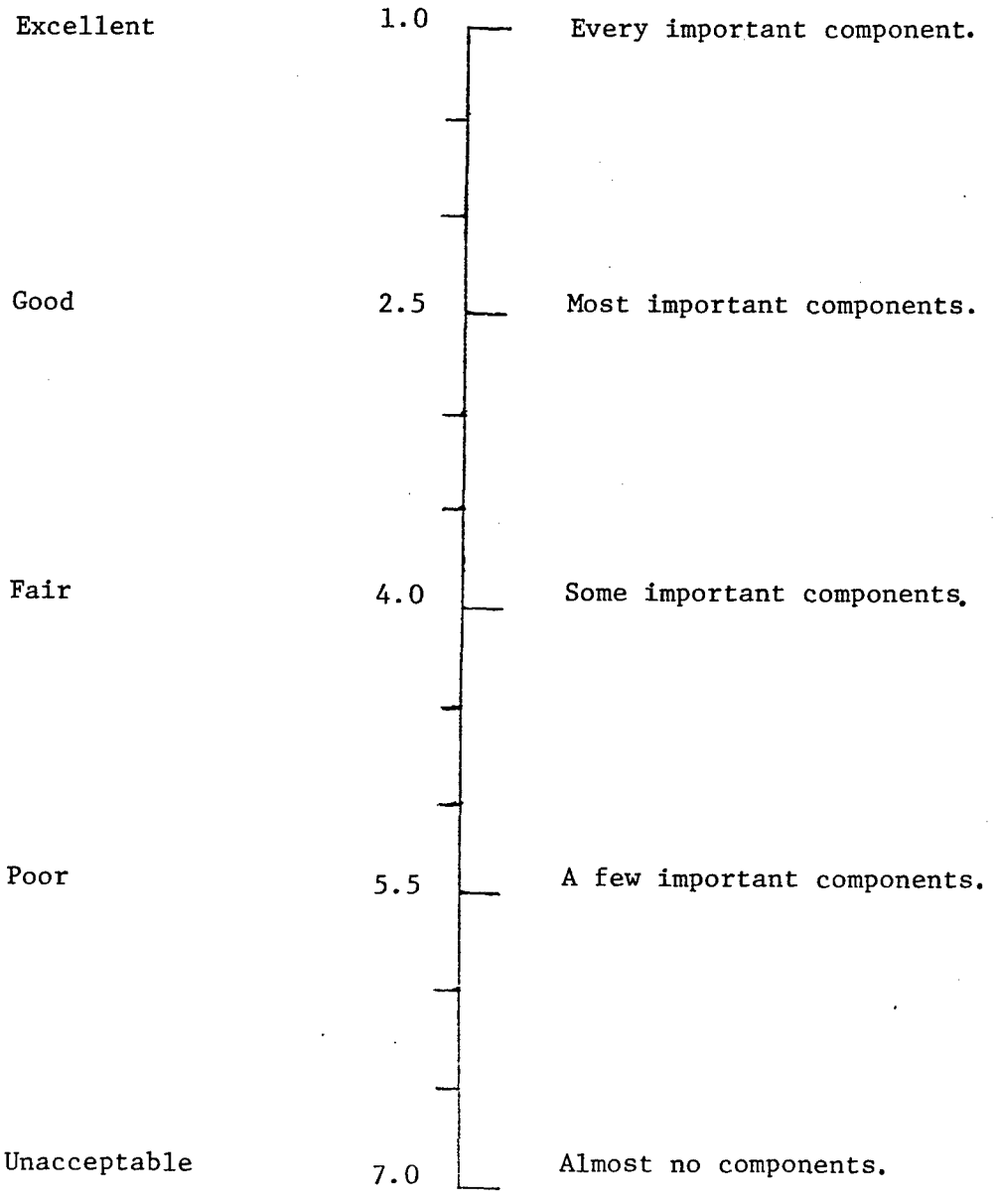
If rating is poor or unacceptable, please comment further.

Rationale: Accessibility of internal components is defined by the number of other components one must remove first and (as a consequence of this) by the effort involved in reaching the desired component. It is unlikely that this type of accessibility can be objectively measured without great difficulty.

13. TEST POINT AVAILABILITY

RATING ITEM

Test points are available to check



If rating is poor or unacceptable, please comment further.

Rationale: Test point availability is considered to be a most significant factor affecting the capability to troubleshoot an equipment. There are other factors, such as the accessibility of these test points, that bear on trouble-shooting capability, but none is as important as availability. The dimension represented on the scale is the correspondence between the number of test points and the number of major components that require testing. One could, of course, check this factor out objectively by examination of the equipment design, but we have taken the tack that the test participant, in working with the equipment, is in the best position to know how this correspondence works in actual practice.

14. EASE OF TROUBLESHOOTING MALFUNCTIONING EQUIPMENT

RATING ITEM

The malfunctioning component can usually be discovered with

Rating		Descriptor
Excellent	1.0	Almost no effort, difficulty or time.
Good	2.5	Slight effort, difficulty or time.
Fair	4.0	Moderate effort, difficulty or time.
Poor	5.5	Great effort, difficulty and time.
Unacceptable	7.0	Exceptional effort, difficulty and time.

If rating is poor or unacceptable, please comment further.

Rationale: This scale deals with troubleshooting as a total function. Equipment characteristics (such as accessibility of internal components and test point availability, scales for which were described previously) affect troubleshooting but are not maintenance functions per se.

It is possible to measure the operator's troubleshooting proficiency on the job, but to do so requires that either an observer must measure repair time or the operator himself must report this time. Often this is not feasible in the context of a test operation. Moreover, objective troubleshooting measures do not get at the effort/difficulty dimension represented in this scale (along with time, which these objective measures do deal with). Hence, use of such a scale can provide useful information describing the ease or difficulty of keeping an equipment running.

15. GENERAL EQUIPMENT MAINTAINABILITY

RATING ITEM

Preventive and corrective maintenance can be accomplished

Rating		Descriptor
Excellent	1.0	With almost no effort; very rapidly.
Good	2.5	With slight effort; quickly.
Fair	4.0	With moderate effort; acceptable time.
Poor	5.5	With much difficulty; long time.
Unacceptable	7.0	With excessively strenous effort, difficulty and time.

If rating is poor or unacceptable, please comment further.

Rationale: This scale describes general equipment maintainability, including both preventive and troubleshooting aspects. Consequently, it subsumes the preceding maintainability scales. Because of its generality, however, it is assumed that one would wish to use this scale only in conjunction with one or more of the previous ones. This scale provides a summary quantitative evaluation of maintainability from the technician's standpoint. It is unlikely that such a summary statement could be made objectively except as a conclusion based on a number of empirical tests, which might be difficult to perform in an OST context. The scale dimensions are those most pertinent to maintainability: effort and speed.

16. SAFETY

RATING ITEM

Required safety equipment

Rating		Descriptor
Excellent	1.0	All required safety equipment is available and accessible.
Good	2.5	Almost all required safety equipment is available and accessible.
Fair	4.0	Most required safety equipment is available and accessible.
Poor	5.5	Only certain items of required safety equipment are available and accessible.
Unacceptable	7.0	Very few required items of safety equipment are available and accessible.

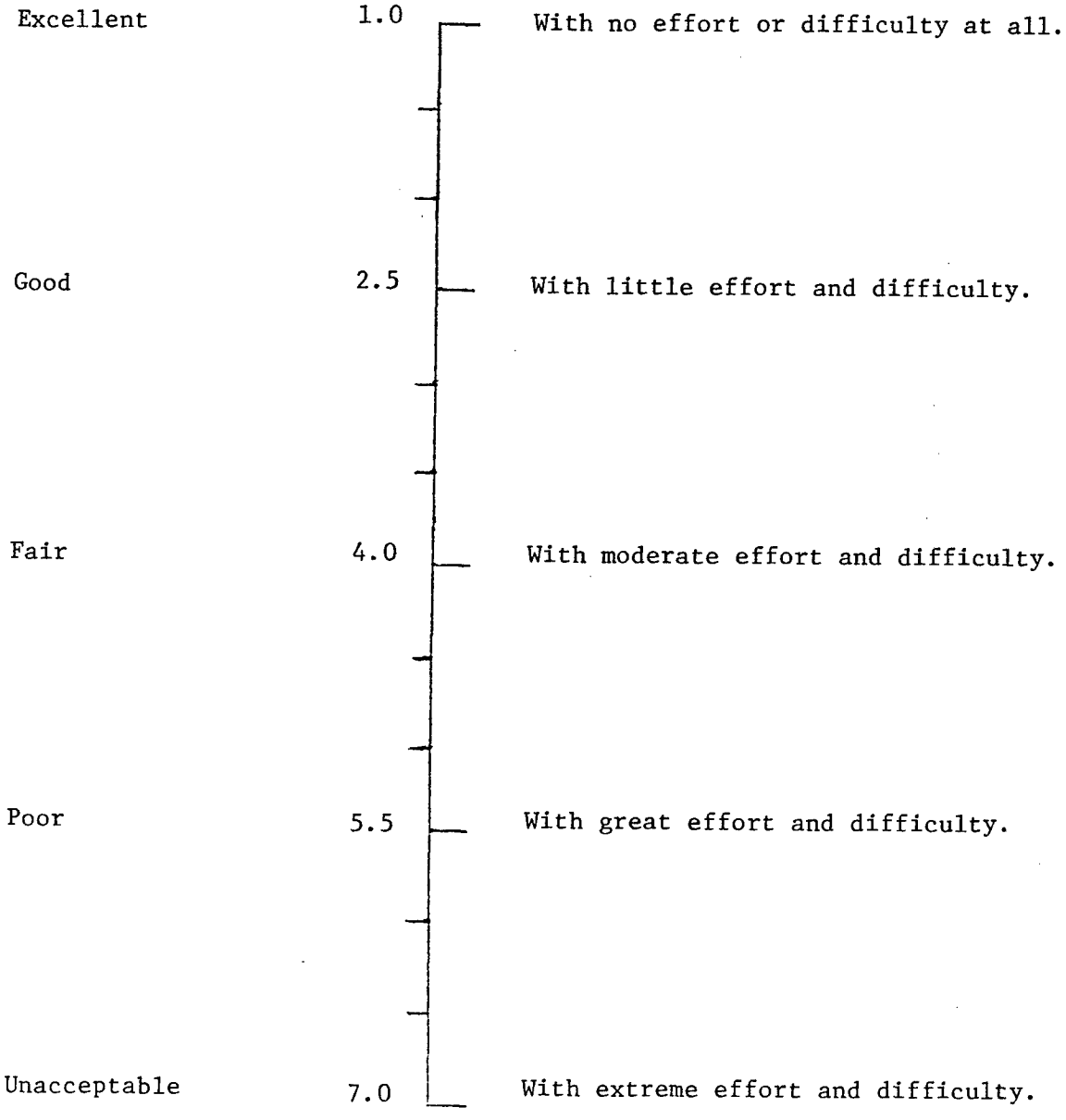
If rating is poor or unacceptable, please comment further.

Rationale: This scale assumes that all required safety equipment must be available and accessible in the vehicle, ground facility, or weapon system. To the extent that less than all such equipment is available, the system is deficient.

17. OPERATING PROCEDURES AND/OR TECHNICAL MANUALS

RATING ITEM

(Operating Procedures and/or technical manuals)* can be understood and followed



If rating is poor or unacceptable, please comment further.

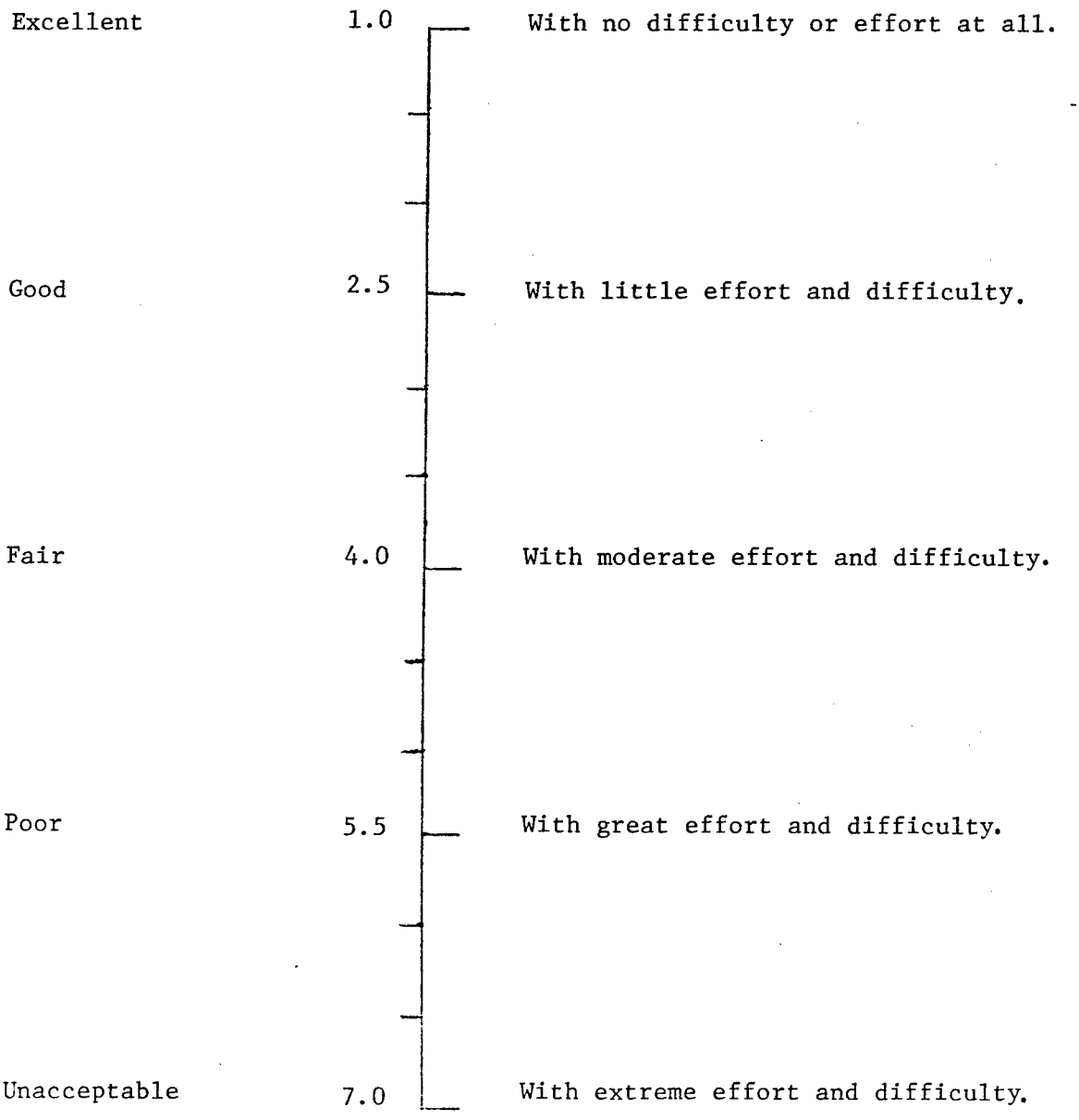
* Select one

Rationale: A major factor affecting how well personnel perform their jobs is their ability to understand and follow the procedures and technical documentation they must employ. This scale is designed to measure the operator's evaluation of this factor. Errors in performing procedures can of course be measured objectively, but do not describe the effort involved in using procedures and technical manuals. It is therefore almost impossible to evaluate procedures and technical manuals objectively; i.e., without securing the operator's opinion on the matter. As usual, the effort factor is the scale dimension.

18. WORKLOAD

RATING ITEM

My job can be performed effectively

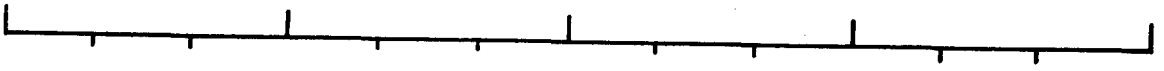


If rating is poor or unacceptable, please comment further.

Rationale: It is often of interest to determine just how easy or difficult it is for test personnel to operate their vehicle or equipment. This scale is designed to measure the effort/difficulty associated with that operation. Since the evaluation of this factor is largely subjective, no objective measurement is feasible, particularly within the context of OST. Note that the scale deals with effective performance and the difficulty associated with effective performance. It is assumed that the respondent knows what comprises effective performance.

An alternative way of getting at workload is in terms of the following scale, which is particular to a specific equipment:

Operation (or maintenance)* of the _____ (fill in) equipment is



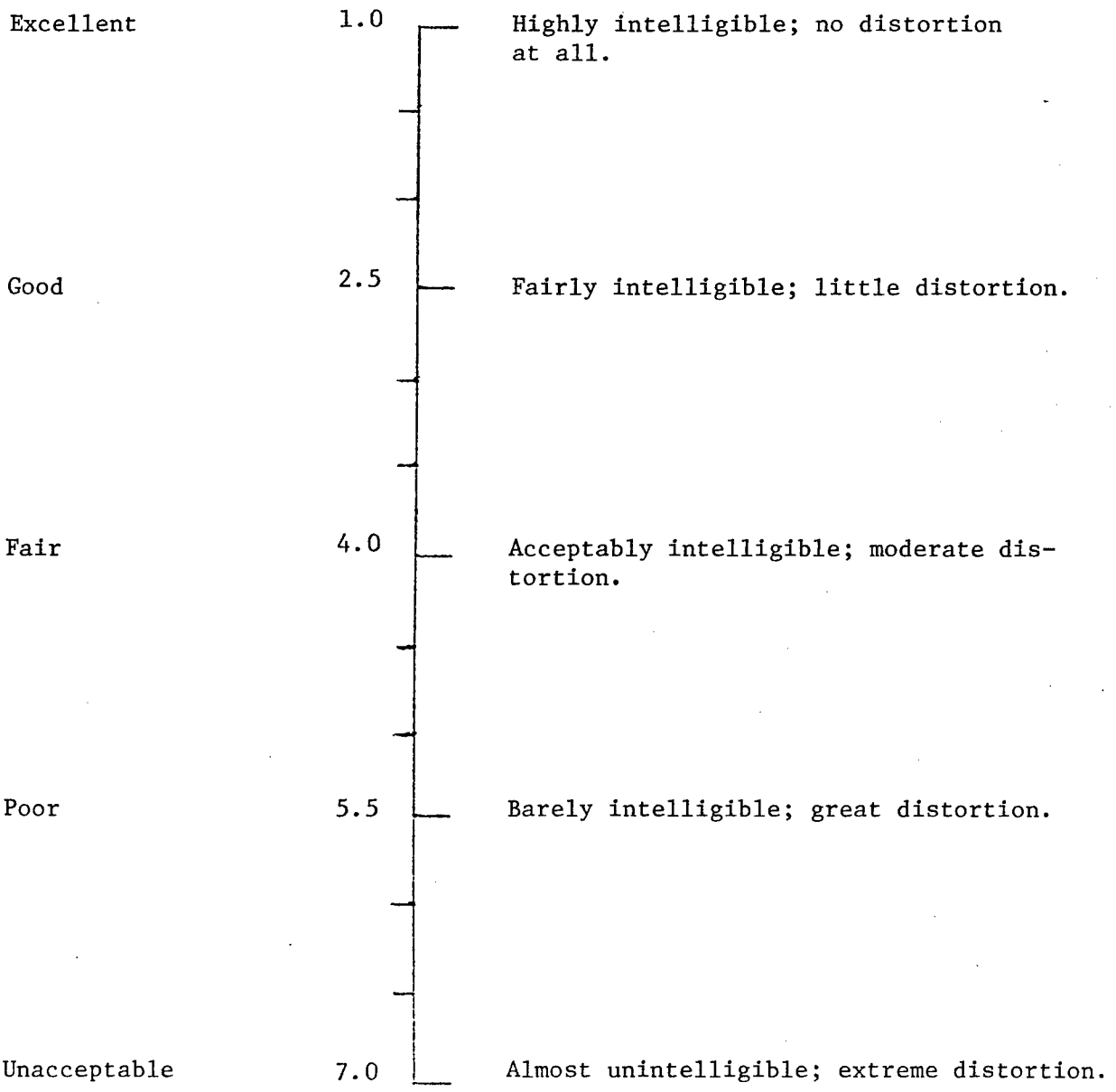
Very Easy	Easy	Neither Easy nor Difficult	Difficult	Extremely Difficult

*Select one

19. COMMUNICATIONS

RATING ITEM

Intercom/radio messages between personnel are



If rating is poor or unacceptable, please comment further.

Rationale: The dimension in this scale is the physical distortion of the communicated message, as perceived by the recipient of the message. Physical distortion of the signal can be ascertained by objective methods, but this may be difficult in the OST context.

CRITICAL INCIDENT REPORT

Please report any incident during the operation of the vehicle, equipment, or weapon system that resulted or could have resulted in an abnormal or potentially dangerous situation (this includes any equipment malfunction).

Check the stage of the operation in which the critical condition occurred and then describe it in as much detail as you think necessary.

Stage of Operation — Subsystem Involved

(This listing must be provided by the evaluator; it is specific to the equipment being evaluated.)

DESCRIPTION

Symptoms (How did you first notice this problem?)

Diagnosis (How did you determine what the problem was?)

Causes (What produced the problem?)

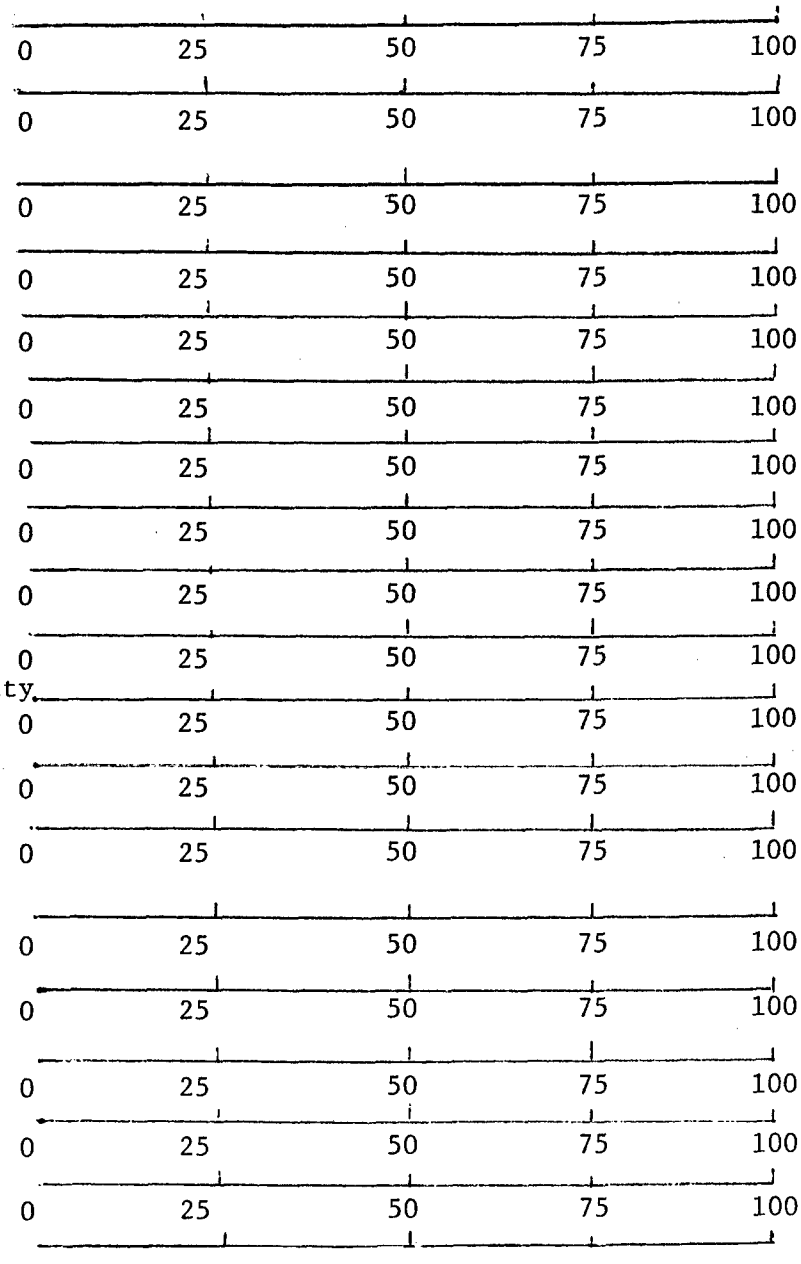
Remedy (What did you do to solve the problem?)

Rationale: Any untoward incident, event, or phenomenon occurring during the test operation may indicate a deficiency that needs improvement before the system is released to operational use. This report is completed at the conclusion of a test run or operation--but only if an incident worthy of note has occurred. Note that this report can also be used to report equipment malfunctions (thus combining data on both personnel and equipment factors).

SATISFACTION CHECKLIST

On a scale from 0 to 100, where 0 represents complete dissatisfaction and 100 represents complete satisfaction, please check your degree of satisfaction with the following factors:

1. Noise, temperature/humidity, vibration
2. Illumination
3. Vehicle handling/driving qualities
4. Display readability
5. Control accessibility
6. Control-display arrangement
7. Information understandability
8. Workspace
9. External visibility
10. Vehicle entrance/exit
11. Internal component accessibility
12. Test point availability
13. Ease of troubleshooting
14. Equipment maintainability, general
15. Communications
16. Safety
17. Operating procedures
18. Technical manuals
19. Operator workload



Rationale: This scale can be used as a very abbreviated summary of the preceding individual scales but obviously does not provide as much information as the latter. It does not indicate the reason for the respondent's satisfaction/dissatisfaction.

SECTION SIX--INTERVIEW QUESTIONS

This section presents questions that evaluators can use in interviewing OST participants to secure data on the adequacy (from a personnel standpoint) of the equipment/system under test. The questions are of two types: general and specific. The interview begins with general questions to determine whether any personnel performance problems have been noted by the operator/maintainer. More specific questions follow to cover the range of equipment/job variables that could influence performance.

- -

The questions listed in this section are to be asked in an interview situation. It is assumed that if the Operational System Test consists of several test operations or cycles, the test operator (or maintainer) will be interviewed following each such operation/cycle.

These questions cover the most important topics that describe how test personnel operate and maintain the system under test. The following topics are covered:

1. General questions.
2. Equipment characteristics.
3. Environment.
4. Job aids.
5. Safety.
6. Manning.
7. Training.
8. Information.
9. Communications.
10. Maintenance.

These topics are roughly the same as those covered in previous sections, but have been created specifically to be used as part of an interview.

In all cases the interviewer should begin his interview with the general questions, because these permit the interviewee (respondent) to produce information that he considers most relevant and important.

If the respondent produces significant material as a result of the general questions, he should be allowed to direct the interview into channels he wishes to follow--as long as these channels are relevant and, in the opinion of the interviewer, productive.

The more specific questions should be asked when the respondent has exhausted the material he produces as a result of the general questions or if he has not touched on any of these topics previously. They should be asked if the respondent appears uncertain as to how to proceed, is not producing any information of value, or is generally noncommunicative.

If several test operations are performed and the operator has been interviewed following each test operation, it is permissible to ignore those questions that have been asked previously and for which the test situation has not changed so that previous answers are still applicable (but not if the operator has been learning on successive operations). Under these circumstances, only the general questions need be asked.

Although it is unnecessary for the interviewer to follow the precise wording of these questions or the order in which they are presented here, the general tenor of the questions should not be changed.

The emphasis in these questions is on problems or difficulties experienced by the operator/maintainer, rather than on how well these personnel have performed. These difficulties indicate inadequacies in the equipment, procedures, etc., that should be cleared up before the system is released to general use.

GENERAL QUESTIONS

1. Can you recall any difficulty or problem (no matter how small) you experienced during the previous test operation(s)? If so, what was that difficulty? What do you think was responsible for it? What actions did you take (i.e., what did you do about it)?

Comment

It is always best to begin an interview with a general all-encompassing question such as the above. This permits the respondent to focus on what appears to him to be the most important of the difficulties he has experienced.

2. Did you observe anyone else in your team having any difficulty? If so, what was the difficulty? Do you know why he had this difficulty? What action did this individual take to relieve the difficulty?

Comment

This question is asked only if the operation is a team affair. All personnel on the team should be interviewed to cross-check individual answers.

3. Were there any characteristics of the job, such as the equipment, procedures, technical manuals, tools, weapons, etc., that made it difficult for you to do your job? How do you know that you had more than your usual difficulty?

Comment

Again, a general queston that allows the respondent to select what he considers the most important topic on which to zero in. It is possible that the respondent in answering question 1 will also discuss equipment/job characteristics that presented difficulty; but this question should be asked in any event, since it directs the interview to the equipment/job as a whole. It is desirable also to return to this question later in the interview in the following manner: We have discussed certain characteristics of the equipment/job that were not optimal. Can you think of <u>any other</u> characteristic that bothered you?

4. Assume that someone less skilled than you had to do your job. Would there be anything about the equipment, procedures, or the job as a whole that would cause a less skilled man difficulties?

Comment

It is assumed that the operator being interviewed is reasonably skilled. (Information about the type and level of training the interviewee has received should be ascertained prior to this interview.) However, it is entirely possible (even likely) when the equipment enters the operational inventory that someone less skilled will perform the job. An equipment/job feature that presents no difficulty to a skilled man may, however, have entirely different effects on a less skilled man. Test personnel can often estimate what these effects would be, and it is important to know these effects because if they are severe enough, some equipment or procedural change might be necessary. In asking this question, it may be necessary to provide an explanation of the reason for asking it, as described in the Comment.

5. Did the equipment fail in any way to perform as it was supposed to do (in any respect, no matter how small)? If so, do you know why? What did you do in reaction?

Comment

Any deviation ought to be routinely reported and might well be, but it is desirable to remind the interviewee about this possibility.

6. Did the test operation as a whole, or your job during the test, take much longer than you had expected it to take? How much longer? Do you know why? Did this extra time affect your performance in any way?

Comment

If the test operation took substantially longer than was planned, this indicates that something is wrong and the operating procedure may have to be changed. In any event, the evaluator must find out if the extra time affected the operator's performance.

7. Can you think of any changes in equipment, procedures, or the job as a whole that must be made to accomplish the mission? That should be made, if at all possible? Or that could improve the equipment, procedures, or job as a whole?

Comment

Test personnel often have excellent suggestions for improving their job (e.g., simplifying it, making it easier, or more accurate) and these should be elicited. Asking this question, moreover, cues the interviewee to think of deficiencies in equipment operation that might not previously have been reported.

8. Could you operate the equipment in accordance with the procedures you had been taught (that are in the manual)? If not, did you add any steps? Delete any steps? Perform one or more steps differently? What were these steps and why did you make these changes?

Comment

The evaluator should routinely ask about any procedure changes that were required because these changes should be included in revised technical manuals and might indicate a problem in the equipment/job.

9. In your opinion, would personnel in the field have any difficulty in operating/maintaining the equipment?

Comment

Assuming that the test personnel have worked in the Fleet Marine Force prior to becoming test personnel, they will remember their past experience.

10. Is every part of the equipment necessary?

Comment

Test personnel may have found no use for a designed feature that was considered essential during design. There may be valid justification for elimination of costly "extras."

11. Did you experience any difficulty as a consequence of operating the equipment for a prolonged period of time?

Comment

Equipment should be evaluated for a full shift and longer. Field or combat operations often require extended operations.

EQUIPMENT CHARACTERISTICS

In general, questions concerning equipment characteristics should follow general questions, although, if difficulties described in relation to general questions deal with some job aspect other than controls and displays, the interview should logically proceed on those other aspects, returning to controls and displays later.

Note. If one receives a negative answer to any of the following items (i.e., there is no difficulty), the interviewer would not proceed with the follow-up questions included in that item.

1. Were the controls difficult to operate? Any particular controls? Do you know why the controls are difficult to operate? How does the difficulty show itself (e.g., mushiness, sluggishness)? What was the effect of the difficulty on your performance?

Comment

If controls are difficult to operate, the effect on personnel performance is likely to be severe, particularly if the controls are in a vehicle. The interviewer should ask whether the operator knows why the difficulty exists, even though the operator often is unable to answer. The symptoms of the difficulty and the effect of the difficulty on the operator's performance may serve as clues to diagnosing the problem and determining its importance.

2. Were any of the controls difficult to reach? Which ones? How important are these controls? What effect does this difficulty have on your performance?

Comment

This question deals with control accessibility. It is necessary to determine the importance of the inaccessible controls and the effect of their inaccessibility on performance to properly evaluate the significance of the problem raised.

3. Were any of the displays (e.g., meters, indicators) difficult to read? Which ones? How important are these displays? Why were they difficult to read? What was the effect on your performance?

Comment

As in question 2, the interviewer seeks to determine which displays are difficult to read, how important these are, and the effect on performance, because these indicate just how important the problem is.

4. Did the displays provide all the information needed to do the job? What information was missing? Was there too much information? Was any of the information unnecessary or irrelevant to the job? How important was this factor in affecting your performance?

Comment

Information presented via displays may be too much, too little, unnecessary, or irrelevant. This question seeks to distinguish among these possibilities. The specific content of the information presenting the problem should be ascertained.

5. Were any of the displays difficult to understand? What precisely about the displays was difficult to understand? Which displays? How did this difficulty affect your performance?

Comment

Interviewees may have problems describing precisely what is meant by understanding. What we refer to here is the interpretation of the meaning of the information presented.

6. Did you have any difficulty reading the lettering or indicator lights when they were unlit? Lit? Which ones?

Comment

Many indicators are identified by labels which must be read even though the indicator is unlit.

7. Did any of the controls or displays seem unnecessary to perform the job? Which ones?

Comment

Ordinarily one would expect every control/display on the equipment or in the vehicle to be needed to perform the task. Occasionally, however, the nature of the job changes between the original design and the OST. This question enables the interviewer to check on this factor.

ENVIRONMENT

1. Was the lighting in your area (or vehicle) inadequate at any time for you to operate with maximum efficiency? Too little lighting? Too much lighting (glare)? What was the effect of this on your job performance?

2. Was any area (or any part of the vehicle) in which you worked excessively noisy, improperly ventilated, too cool, or too warm? How did this affect your performance?

3. Was there too much vibration in the vehicle when it was driven? Did this affect your work? In what way? How much?

4. Was there insufficient room around the equipment you operated so that it was difficult to move about? Was this true of the equipment in general or of a specific equipment only? What were the effects of this on your job? How great?

5. How difficult is it to get in and out of your vehicle? Does this affect your job performance? In what way?

Comment

These questions are obvious and consequently specific comments are not made about individual items. The operator's working area (i.e., his environment) may affect the efficiency with which he performs his job and so the evaluator will wish to examine the various aspects of that environment. Particular emphasis should be placed on the effect of these environmental factors on performance. If a factor has little effect on job performance, it will be unnecessary to probe deeper.

1. Are the tools and equipment you would need for maintenance available? Appropriate? Satisfactory? Are there any special tools you might need that are not available to you? If so, which ones?

2. Are all authorized spare parts available? Were any spare parts required and not available?

Comment

Tools and spare parts for maintenance fall into the category of job aids. If a problem in relation to these arises, it is probably because some of these tools/spares are either inappropriate or missing. The evaluator will also wish to know whether any tools, equipment, or spare parts are required that were not anticipated during design.

SAFETY

1. Is there any safety equipment you need that has not been provided?

2. Are there any desirable safety features (e.g., interlocks), that have not been included in the design of the equipment you operate? What are these features? How important are they?

3. Are there any safety hazards in the vehicle or area that you noticed? If so, what are they? Is all safety information conspicuously posted? Can anything be done to make it easier for the operator to heed these warnings?

4. Are all required safety equipment available and accessible in your area (vehicle)? If not, what is missing? Is there any safety equipment you need that has not been provided?

1. Could you have used more men to do the job than were assigned to your team? If so, how many and of what type and what skill level should they have been? Could you have used fewer men to do the job? If so, which ones would you eliminate?

2. Was anyone on your team overloaded? Excessively fatigued by the end of the test operation? Why? What effect did this have on overall performance?

Comment

The reason for asking these questions is to verify that the appropriate number of personnel have been assigned to perform the job of operating or maintaining the test vehicle, weapon, etc. The question on workload seeks to determine indirectly whether more personnel are needed (if anyone is overloaded, presumably he needs help to carry the load).

1. Do you feel that the training you were given for this job was appropriate? Inappropriate? In what ways? What would you recommend to improve the training?

2. Are the men in your team properly qualified in terms of training?

3. What items were missing from the training you received that should be added? Was there anything about the training you received which you considered unnecessary or which you did not understand? Did you receive enough training to do the job?

4. What parts of the training were most important for safe, efficient operation? What parts were least important?

Comment

The OST is the first opportunity the Marine Corps has to check on the adequacy of the projected training that will be given personnel to operate the system. Test personnel will have been given factory training, but until OST the opportunity to check the adequacy of the training against performance has been lacking. The above items seek to gather information on the adequacy of training plans.

1. Do you feel that the procedure for operating the equipment (system) is completely adequate? Does it reflect what you have to do? Does it cover all contingencies? What is missing? What is included that is unnecessary? How could it be improved?

2. Are all required TO's, handbooks, etc., available to you? Are they complete; i.e., do they cover everything you need to know about the equipment? What was missing from these? Was any unnecessary material included? Is the material understandable? How could these publications be improved?

3. Have you had occasion to refer to technical manuals since you began the OST? On what occasion? To find out what?

Comment

Actual operating procedures may differ somewhat from those that were developed during design. These questions are asked to elicit any required procedural changes. Technical manuals should reflect the needs of the operator and, like operating procedures, may have to be brought up to date.

COMMUNICATIONS

1. Did you have any difficulty in receiving or supplying information to other personnel over internal communications equipment? What were the causes of this difficulty? How can these be changed?

2. Did the necessity for communicating interfere in any way with your job of operating the equipment? To what extent?

3. Did you have any difficulty in providing required information? Why?

MAINTENANCE

General

The following questions are asked only once (at the conclusion of the OST) and refer to the interviewee's total experience in performing maintenance:

1. Have you had any difficulties or problems in performing preventive maintenance (cleaning, oiling, adjusting, etc.) on the equipment? Did anyone else on your team have these difficulties? What were these difficulties? What caused them? How significant were these difficulties? How could these problems be eliminated?

2. Have you had any difficulties or problems in performing corrective maintenance during the test operation? What were these difficulties? Did anyone else on your team have the same difficulties? What caused these difficulties? How significant were they? How could they be eliminated?

3. How often has it been necessary for you to perform corrective maintenance during the OST? What piece of equipment failed most frequently? What impact did this have on test operations?

4. Was there anything about the equipment, procedures, or the tools you used that might make it difficult for persons less skilled than yourself to troubleshoot your equipment? Which equipment? What procedure? Which tools? What caused the difficulty?

Specific

The following questions are asked following each test operation:

1. Did you have an equipment malfunction during the test? (If the answer is no, the following questions need not be asked.)

2. How did you first become aware (by what displays or other symptoms) that a malfunction had occurred? What were the symptoms?

3. When the malfunction occurred did you have enough information to know what caused it?

4. Did you try to troubleshoot the equipment (bring it back on line) during the test? If not, why?

(The following questions need be asked only if the interviewee has attempted to troubleshoot the malfunctioning equipment.)

5. How easy or difficult was the malfunction cause to diagnose?

6. Did you refer to your technical manual? Was it of any value? How useful do you find your TM generally in troubleshooting?

7. Did you have any difficulty in securing access to the inside of the equipment? In unfastening panels? Removing the equipment chassis?

8. Approximately (to the nearest minute) how long did it take you to determine the cause of the failure? To remove and replace a component? To check that the equipment was working again? Do you consider this time excessively long? Average? Short?

9. Did you have enough room to move around the outside of the equipment while you were troubleshooting it? Within the equipment?

10. Did you have all the proper tools to perform the maintenance? Any missing? Which ones were missing? Which ones were inappropriate?

11. Were there enough test points to check out the equipment? Did you have any difficulty finding them? Are they located close to the units they check? Are they accessible?

12. Did you have a spare component to replace the failed one? Did you have to go elsewhere to find a spare?

13. Were there any safety hazards in troubleshooting the equipment? What were these? What caused them? What could be done to eliminate them?

14. Were there any difficulties in removing the failed component because of weight, shape, location?

15. Was there any difficulty installing the replacement unit?

16. Did you ask anyone else's advice while troubleshooting? Did you work as part of a team in repairing the failure?

17. Was the failure successfully cleared up? If not, what happened then?

Comment

The intent of these questions is to secure as much information as possible about maintenance, and particularly about troubleshooting equipment that has failed during the test. The questions in some respects parallel those asked about other aspects of the system (e.g., information, job aids, safety) but they must be asked again in relation to troubleshooting because their significance is different in maintenance.

SECTION SEVEN--TEST PROCEDURES

The preceding sections have dealt largely with test planning. This section describes procedures to be followed during the actual conduct of the personnel performance test.

- -

In order to secure meaningful results from the Operational System Test, the following are required:

1. Scenario Deviations

Test personnel must follow the scenario (operations plan) laid out for them. If the test objective is to determine the time for a tank to drive 1 mile, ford a 2-foot deep creek, and then drive 5 miles, this scenario must be followed in one complete and continuous action rather than as separate parts over different periods or on different days.

2. Briefing Test Participants

To ensure that the above is performed, ascertain before the test begins that the test participants have all the equipment, operating procedures, performance aids, etc. that they need to do their job properly. Test participants should be briefed before they begin the test operation to ensure that they know what they are supposed to do. This includes not only the route to be followed (if a vehicle is to be driven), but also any information they are supposed to supply to data collectors. Ask them if they have any questions and answer these.

3. Noninterference

Once a test operation has begun, no one should interfere with the performance of that operation by (a) aborting the test or (b) providing information to test participants. The reason for this is that the evaluator is trying to replicate operational conditions and outside interference or aid will not be available to personnel in the operational situation. Any such interference and aid merely cause the test results to be nonrepresentative and nongeneralizable to the operational situation.

The only exceptions to this rule are if (a) a dangerous situation arises that could hazard personnel or the system under test or (b) equipment malfunctions and makes the continuation of the test impossible. Data collectors/observers should have the authority to call off a test, but only for the most pressing reasons. Under all other circumstances, data collectors/observers should not interfere once the test has begun. In fact, observers in the physical proximity of test participants should be as unobtrusive as possible and should provide no assistance, even when asked for it, except under the hazardous conditions referred to previously.

4. Unforeseen Occurrences

However, measurement in a field environment always involves the possibility that the unforeseen will occur. For example, equipment or instrumentation may fail or the weather may not be appropriate for a particular test. The Test Conductor should have contingency plans in the event that a change in the test plan is required. Such a contingency plan may involve rescheduling a test operation, or performing part of rather than the entire scheduled event. Observers should be made aware of these contingency plans.

5. Reasons for Deviations

If test personnel fail to follow the scenario exactly, it is the observer's responsibility to determine why this has occurred. However, he should not interfere with the deviation while it is occurring, unless it involves a hazard situation, nor should he call the test participants' attention to the fact that they are deviating from the scenario. Following the test operation, he should question test participants to determine why they deviated, because the reason may have some bearing on the adequacy of the system and the procedures developed for it.

6. Specific Test Objectives

Each test operation should be defined in terms of the specific test objective it is designed to satisfy. The Test Conductor should use a check-off sheet to record that each test has in fact been performed fully, partially, or not at all. This is particularly important when evaluating a system of any size or complexity.

7. Manual Backup to Instrumentation

If data will be collected by means of instrumentation, the Test Conductor should have a manual backup method of collecting data in the event that the instrumentation fails.

8. Practice Runs

Prior to the start of formal data collection, at least one or two practice runs following all procedures exactly as intended should be conducted to try out data collection methods. Data collectors should be debriefed following these runs to determine whether any last minute changes in test procedures and/or data collection forms are necessary. Debriefing should focus on whether the desired data can be collected efficiently and whether serious data difficulties are being encountered.

9. Observer Stations

Observers of the test operation should be stationed in such a position that they can see what is occurring without their intruding unduly upon the privacy of test participants. It must be emphasized to data collectors that they are not test participants.

10. Equipment Failures

Data collectors should record all instances of unscheduled events occurring during the test operation. The most important of these will probably be equipment and logistics failures and any repair activities performed. This information may be needed to explain performance results.

11. Data Collection Forms

All data collection forms should be controlled in terms of their issuance to data collectors from a central office; they should be returned to the same office. All data collection forms should, as a minimum, contain the following information:

a. Identification (e.g., name) of the test participant.

b. Name of the data collector.

c. Identification of the test operation for which data are being collected.

d. Identification of the equipment being tested (in case this is not implicit in the name of the test operation).

e. Date the test was performed.

f. Scenario number.

12. Data Quality

It may be useful for data collectors to record their judgment of the quality of the data being collected in a particular test operation when, for whatever reason, they have little confidence in those data. Such information would be useful to the Test Conductor in drawing conclusions from the data.

13. Interviews

Test participants should be interviewed following each major test operation. They may be able to supply information which would amplify and explain observers' data.

14. Data Collection Monitoring

The Test Conductor should monitor all data collection activities on a sampling basis. This is to ensure that his personnel are performing as desired and that he will secure the data he desires.

15. Start/Stop Time

The start and stop time of the test activity being monitored should be ascertained by the data collector.

16. Reliability Data

All equipment failures observed to occur by test participants should be recorded by data collectors. The following information should be collected:

a. Time the failure was observed by test participants.

b. Symptoms of the failure.

c. Any diagnostic, troubleshooting, or repair activities performed by test participants.

d. Time the equipment was restored to operating status.

e. Whether or not the test was aborted as a result of failure.

f. How serious the failure was in terms of its impact on the accuracy and precision of the test operation.

One of the major parameters in terms of whether or not the system under test will be judged effective is system reliability, as measured by the occurrence of equipment failures. It is therefore essential that all such failures be reported, no matter how trivial they may appear to be on the surface. Since most of these failures will occur or first be noted during a test operation, both test participants and data collectors should be admonished to report them.

17. Availability

Availability is another important system parameter. Essentially, availability is a measure of the extent to which the system is ready to perform when it is needed.

$$\text{Availability} = \frac{\text{Total uptime (system actually operating)}}{\text{Total time (system in usable condition)}}$$

Obviously, any failure may cause the equipment or system to "go down"; until it is restored, that equipment or system is not available for use. Not every equipment malfunction will necessarily cause the equipment or the total system to fail (e.g., a light on a console failing); but every failure must be reported, if not for the determination of availability, then for the determination of reliability.

18. Maintainability

The length of time it takes to restore an equipment to operational status (otherwise known as the mean time to repair) is one index of the maintainability of the equipment or system. Other indices of maintainability relate to the equipment characteristics that make it easy or difficult to troubleshoot the equipment (e.g., accessibility of components, availability of test points, etc.). Both should be reported.

It should be obvious that the key to these measures--reliability, availability, and maintainability--is the failure report. Hence data collectors should make every effort to report the details of such failures, preferably on special report forms designed for this purpose.

SECTION EIGHT--INTRODUCTION TO STATISTICAL METHODOLOGY

Introduction

Researchers, investigators, and policy makers are often faced with the problem of obtaining or evaluating data relevant to the solution of a specific problem. Data obtained and analyzed using proper statistical techniques are likely to yield knowledge vital to the understanding of complex problems while data improperly obtained or analyzed will frequently result in poor understanding or erroneous conclusions.

In essence, statistical methodology is concerned with planning and carrying out the collection, tabulation, and analysis of data. Statistical methodology may be subdivided into two broad areas--descriptive statistics and statistical inference. Descriptive statistics is concerned with the development and utilization of appropriate arithmetic, tabular, and graphical techniques for describing data in an orderly and meaningful way. Statistical inference describes the methodology for making statements that go beyond the data that have been observed or analyzed. This chapter provides guidelines for the use of appropriate statistical techniques.

Statement of the Problem

The application of statistical methodology should not be undertaken without a clear statement of the problem being investigated. A well-defined problem statement should include clear definitions of:

1. The population(s) or universe(s) under study. The totality of individuals or units about whom knowledge is desired must be clearly specified.

2. The aspect(s) or the population(s) of interest. The characteristic(s) of the population(s) that are being studied must be rigorously defined.

3. The purpose or goal of the research. The primary goal and its associated objectives must be clearly stated. Objective(s) may include the estimation of unknown values, the answer to a specific question about a population, a comparison between populations, or the investigation of a relationship between various aspects of a population.

Some examples of simple problem statements are as follows:

1. Example 1. A new tire has been developed but it is not clear how it will perform over rough terrain. Since performance is characterized by the tire's tread life (in miles), it is necessary to estimate the average number of miles over rough terrain that the tire will travel until it needs to be replaced.

   a. Population: All new tires of this type.
   b. Aspect: Tread life (in miles).
   c. Purpose: Estimation of average tread life (in miles) over rough terrain in order to gauge tire quality.

2. Example 2. A training manual has been developed that should enable new marine recruits to utilize a specific piece of machinery with no further instruction. It is necessary to determine whether this manual is, in fact, effective so that classroom time can be diverted to other essentials.

a. Population: All new marine recruits.
b. Aspect: Ability to use a piece of machinery after reading manual.
c. Purpose: To establish whether manual is effective so that classroom time can be used for other purposes.[1]

3. Example 3. A new reading course has been established and it is not clear whether its effect will be the same in two different areas of the country. The intent of the course is to increase the reading comprehension of poor students.

a. Population: I--All poor students in area I; II--All poor students in area II.
b. Aspect: Improvement in reading score after exposure to a specific course.
c. Purpose: To determine whether the new reading course will improve reading comprehension by the same amount in two distinct areas.

4. Example 4. A final exam is given to all individuals who enroll in a given computer programming course. It is desired to use a qualifying test to determine whether an individual should be admitted to the course, but the relationship between the qualifying test and course performance is not clear. Therefore, an investigation of the relationship between final exam grade and qualifying exam grade is initiated.

a. Population: All individuals who might enroll in computer course.
b. Aspect(s): Qualifying exam score and final exam score.
c. Purpose: To determine whether the screening test is useful; that is, to assess the relationship between qualifying exam score and final exam score.

Variables

Once a statistical problem has been clearly defined, it is necessary to obtain and utilize data pertinent to its resolution. Data analyzed for statistical analysis are usually obtained from either multiple physical or mental measurements or responses to a questionnaire. It is essential that the measurements analyzed be obtained under conditions relevant to the problem being addressed. For example, if tread life of tires over rough terrain is of interest, tires should be tested under conditions that are analogous to the type of terrain about which inferences are to be drawn.

---

[1]The word "effective" was not defined precisely. In order to utilize statistical methodology, the criteria determining "effective" must be clearly stated. As an example, an "effective" manual might be one such that at least 85 percent of all recruits would properly utilize the machinery after reading the manual.

A later section of this chapter will discuss sampling techniques for obtaining data. It should be obvious that data should not be analyzed unless they are obtained in a manner such that they represent the population(s) of interest.

Recalling that statistical techniques are utilized to analyze various aspects of populations, it is necessary at this point to introduce the concept of a statistical variable.

A variable may be defined as a characteristic of the population that may differentiate individuals or units within that population. For example, the variable of interest in Example 1 above was the tread life (in miles) of a given tire. Example 2 was concerned with the variable "ability to utilize a specific piece of machinery"; and Example 3, the variable "reading score after exposure to a course." Example 4 investigated two variables: "final exam grade" and "qualifying exam grade." The variables under investigation must be clearly defined prior to the utilization of statistical techniques.

A number of different kinds of variables arise in practice. The selection of the appropriate statistical methodology is dependent upon the type of variable being analyzed. Broadly speaking, there exist two types of variables, quantitative and qualitative.

A quantitative variable is one that is recorded as a numerical value. The variables "tread life in miles," "reading score," "numerical exam grade," "blood pressure," "family income," "number of heads occurring in eight tosses of a coin," etc. are all quantitative variables since they are measured as numbers.

A qualitative variable is one that is not measured in quantitative units. Qualitative variables are defined by specifying a set of two or more categories into which individual population elements may be assigned. We "measure" or "observe" individuals with regard to qualitative variables by assigning each one to a category. Categories should be defined in such a way so that every individual or unit in the population of interest can be classified as a member of one, and only one, of these categories. This is frequently referred to as establishing a set of categories that are exhaustive and mutually exclusive. In Example 2, the variable "ability to utilize a specific piece of machinery" may be considered as a qualitative variable if individuals are rated as either "can" or "cannot." Examples of other qualitative variables include "state of origin," "eye color," "opinion towards a candidate" (will vote for, won't vote for, undecided), etc.

In actuality, there are gradations between quantitative and qualitative variables. One commonly occurring "gray area" is one in which an observation consists of response to an ordered or ranked scale (say, extremely dislike, dislike somewhat, indifferent, like somewhat, extremely like). For a discussion of the theory of measurement, see Ellis (1966) and Churchman and Ratoosh (1959).

Descriptive Statistics

Once data have been obtained, it is frequently necessary to organize and summarize them in a manner so that their meaning can be clearly understood. Descriptive statistics is concerned with the description of data without attempting to draw inferences beyond the individuals or elements from whom the measurements were taken or observed.

Data are usually summarized using tables, graphs, and summary statistics. The necessity for summarizing data is clear, since a mere presentation of observations or measurements (e.g., CAN, CAN, CAN, CANNOT, CAN, CANNOT, CANNOT, etc.) is often confusing and virtually useless. Such presentations provide data but little information about the problem.

Description of Qualitative Variables

Frequency Distributions and Graphs. Data pertaining to qualitative measures of a population are often presented in tables known as frequency distributions. A frequency distribution may be defined as a listing of all possible categories in which the variable values may occur and the number (or percentage) of individuals or units so designated. Referring to Example 2, a frequency distribution of the ability of a group of 200 recruits to use a piece of machinery may appear as shown in Table 8-1.

Table 8-1

Ability to Use New Machinery

Ability to Use Machine	Number	Percent
Could Use Machine	120	60.0
Could Not Use Machine	80	40.0
Total	200	100.0

It is essential that the total number of individuals or units tabulated be specified so as to facilitate proper evaluation of the data. Table 8-1 clearly indicates that the total number of individuals presented is 200.

Qualitative variables are frequently illustrated by means of circle and/or bar graphs. Circle graphs are especially useful when the relative proportion of individuals falling into each category is of interest; and bar graphs, when the absolute number of individuals falling into each category is of interest. Proper procedure for constructing graphical representations of data may be found in Hamburg (1970, Chapter 3).

When more than one qualitative variable is observed on elements in a population, a contingency table is often a convenient method of simultaneously summarizing such data. Consider the following example. Suppose 400

individuals were exposed to one of two teaching methods--200 were assigned to Method I and 200 to Method II. Assume that, at the end of the course, each individual was given an exam that was graded on a pass-fail basis. In this case, observations consist of (1) two measurements per individual, (2) method of instruction, and (3) exam grade. Individual observations might thus consist of pairs, such as Method I, PASS; Method II, PASS; Method I, PASS, etc. A contingency table summarizing these results might appear as illustrated in Table 8-2. Such tables can be constructed with more than two variables.

Table 8-2

Exam Scores for 400 Recruits Exposed to Two Different Teaching Methods

Teaching Method	Exam Grade		
	PASS	FAIL	Total
I	160	40	200
II	80	120	200
Total	240	160	400

Note that all categories of one variable comprise the rows of the table; and all categories of the second variable, the columns. The number in a particular cell of the table, therefore, represents the number of individuals having both a specific teaching method and exam grade. Data presented in this fashion are especially useful when one wishes to analyze the relationship between two variables. A discussion of contingency tables may be found in Neter and Wasserman (1973, Chapter 26).

Summary Statistics. When considering qualitative variables, commonly used summary statistics include the mode and category rankings.

The mode of a frequency distribution of a qualitative variable is defined as the category in which the maximum number of individuals or units have fallen. The mode (or modal value) of the frequency distribution appearing in Table 8-1 is "Could Use Machine." The mode is useful if one wishes to present the specific category that best represents the data being described. The mode should not be used as a summary statistic if two or more categories contain approximately the same number or percentage of individuals. To illustrate, the data in Table 8-3 are bimodal in nature since there are two distinct maximum categories--"extremely favorable" and "extremely unfavorable." Presenting one of the two categories as representative of this data set would be misleading.

Table 8-3

Attitude Towards Tax-Relief Bill
(Based upon a survey of 500 individuals)

Attitude	Percentage
Extremely favorable	30.0
Moderately favorable	10.0
Neutral or Undecided	15.0
Moderately Unfavorable	15.2
Extremely Unfavorable	29.8
Total	100.0

For the individuals represented in Table 8-3, the rank-ordering of the responses (e.g., 1-extremely favorable, 2-extremely unfavorable, 3-moderately unfavorable, etc.) may provide a valuable summary of the data for many applications. A good discussion of descriptions of qualitative variables is found in McCarthy (1957, Chapter 3).

Additionally, when one observes two or more qualitative variables, it is often of interest to measure the association, or relationship, between them. Considering the data of Table 8-2, we might be interested in measuring the relationship between teaching method and grade on exam. (If there is no relationship between two variables, they are known as "independent.") For a discussion of measures of association for qualitative variables, see McCarthy (1957, Chapter 11).

Description of Quantitative Variables

As in the case of qualitative variables, quantitative variables may be summarized in terms of tables, graphs, and summary statistics.

A frequency distribution of a quantitative variable may be defined as a listing of all possible values of the variable and the number (or percentage) of individuals or units within each value. In many practical situations, however, it is not feasible to list all possible values of a variable simply because that number is too large or infinite. For example, tread life (in miles) is a variable whose possible values are limited only by the accuracy of the measuring instrument (e.g., to the nearest mile, tenth of a mile, hundredth of mile, etc.). Similarly, all possible scores on a test (Example 2) may consist of the values 0, 1, 2, . . ., 100. (Note that in the case of a variable such as the number of heads appearing on eight flips of a coin, the values 0, 1, 2, 3, 4, 5, 6, 7, 8 constitute all possible values and may easily be listed.) As such, it is usually necessary to group values of quantitative variables (usually as intervals) when one summarizes values of a quantitative variable. The "best" methods of grouping

values will not be discussed here other than to note that the number of intervals and specific interval values should contain all possible values of the variable and should clarify, rather than obscure, the underlying data. For a discussion of how to construct meaningful intervals, see Yamane (1964, Chapter 2). Tables 8-4 and 8-5 present frequency distributions for variables discussed in Examples 1 and 3.

Table 8-4

Number of Miles Driven Before Tire Failure
(Based upon tests of 200 tires)

Number of Miles of Tread Life	Number of Tires
More than 40,000 miles	10
35,000-39,999 miles	20
30,000-34,999 miles	30
25,000-29,999 miles	40
20,000-24,999 miles	40
15,000-19,999 miles	30
10,000-14,999 miles	20
Less than 10,000 miles	10
Total	200

Table 8-5

Reading Scores of 100 Individuals in Area I

Reading Score	Percentage of Individuals
90-100	10
89-99	20
70-79	50
60-69	5
0-59	15
Total	100

Such data are frequently illustrated by graphs known as histograms and frequency polygons. For a discussion of these graphical techniques see Yamane (1964, Chapter 2).

Note that, by further subdividing the intervals in Table 8-4, we obtain the data appearing in Table 8-6. Finer subdivisions might result in the limiting frequency distribution graphed in Figure 8-1. A possible limiting distribution for the data in Table 8-5 is presented in Figure 8-2. For a treatment of distributions of continuous variables, see McCarthy (1957, Chapter 3).

Table 8-6

Number of Miles Driven Before Tire Failure
(Based upon tests of 200 Tires)

Number of Miles of Tread Life	Number of Tires
More than 50,000 miles	3
40,000-49,999 miles	7
37,500-39,999 miles	10
35,000-37,499 miles	10
32,500-34,999 miles	15
30,000-32,499 miles	15
27,050-29,999 miles	20
25,000-27,499 miles	20
22,500-24,999 miles	20
20,000-22,499 miles	20
17,500-19,999 miles	15
15,000-17,499 miles	15
12,500-14,999 miles	10
10,000-12,499 miles	10
7,500-9,999 miles	7
Less than 7,500 miles	3
Total	200

Quite often, quantitative variables are summarized using various descriptive summary statistics. When one considers summarizing or representing a set of quantitative data, it is natural to search for measures of the "center" of the data and of the "dispersion" or "variability" of the data.

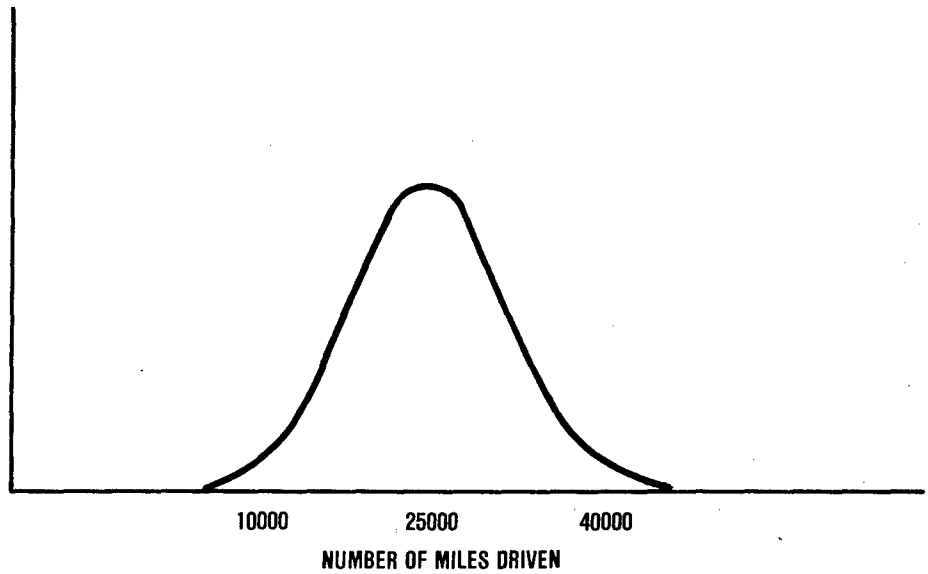


Figure 8-1. Limiting frequency distribution for number of miles driven before tire failure.

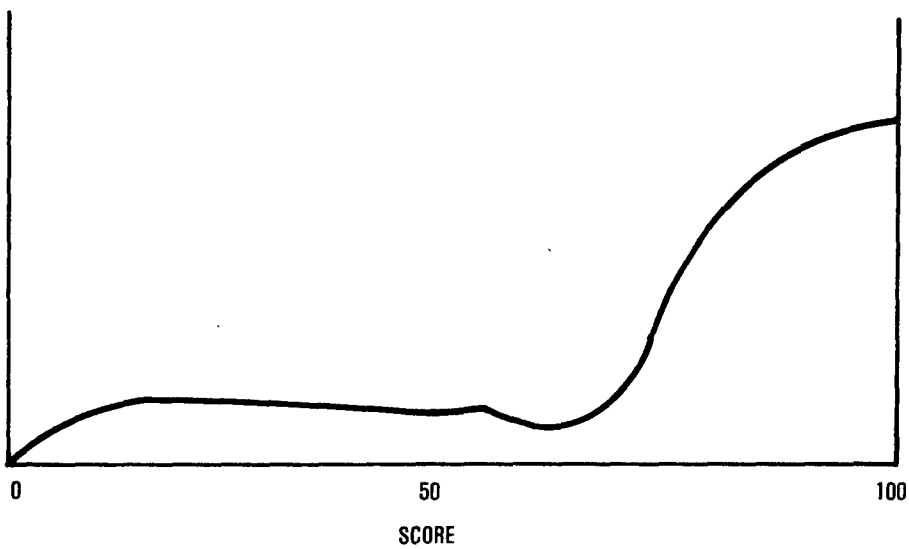


Figure 8-2. Limiting frequency distribution for reading score of individuals in Area I.

Measures of Central Tendency. A variety of measures exist for describing the "center" of a data set.

1. The Arithmetic Mean. The arithmetic mean is defined as the sum of the observations divided by the number of observations. The mean is a useful summary statistic when one wishes to include all observations, including extreme values in the summary measure. Suppose the values 7, 56, 2, 2, 8 represent the number of days during the past 2 months on which a sample of five individuals ate steak for dinner. The mean of the observations 7, 56, 2, 2, 8 is $(7 + 56 + 2 + 2 + 8)/5 = 75/5 = 15$ days. For a more thorough discussion of the arithmetic mean, see Yamane (1964, Chapter 3).

2. The Median. The median of a data set is the value of the middlemost observation (in the case of an even number of observations, the average of the two "middlemost" observations) when the observations are ranked in size order. The median is a useful summary statistic when it is felt that "extreme" observations distort, or are unrepresentative of, the underlying data. For the values listed above (i.e., 7, 56, 2, 2, 8), the median is 7 days since, after ranking these observations (e.g., 2, 2, 7, 8, 56), 7 appears as the middlemost observation. Note that the extreme value, 56, did not enter into the calculation of the median. For a discussion of the median, see Yamane (1964, Chapter 3).

3. The Mode. The mode is obtained in a manner analogous to that of qualitative variables. For a discussion of the mode, see Yamane (1964, Chapter 3).

Many other measures of central tendency exist, such as the harmonic mean, geometric mean, etc. See Yamane (1964, Chapter 3) for a discussion of these and other measures.

Measures of Variability. Although measures of central tendency summarize data in terms of their center, these statistics are in no way descriptive of the dispersion of the data. For example, although Figure 8-3 illustrates two sets of data with similar "centers," the data sets are quite different. If the data sets represent the diameters of a precision tool manufactured by two different machines, the graphs indicate that one machine produces tools of almost uniform quality while the second produces tools with considerably higher variability. Some commonly used measures of variability are discussed below.

1. The Range. The range of a data set is the difference between its highest and lowest values. For the values 7, 56, 2, 2, 8, the range is $56 - 2 = 54$. The range is very sensitive to extreme values, and, like the mean, should not be used when extreme values are felt to be unpresentative of the process or population under study.

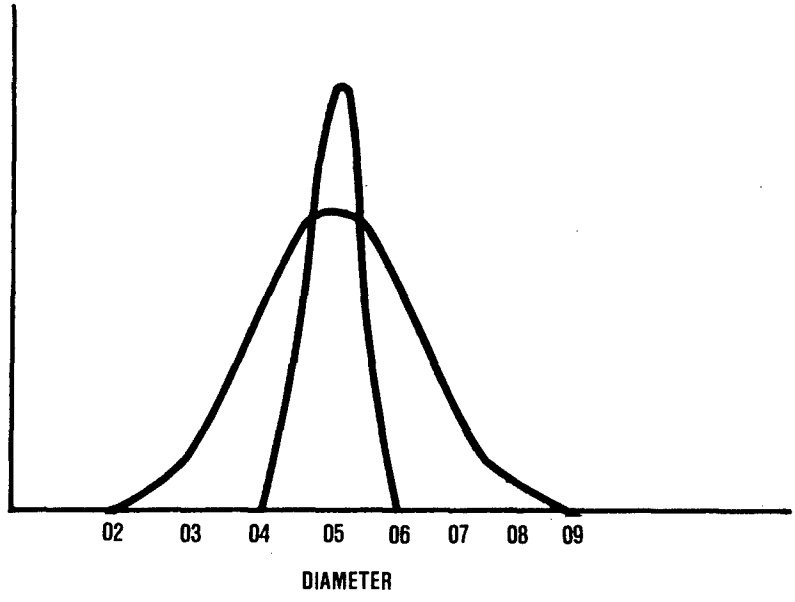


Figure 8-3. Two frequency distributions with similar means but different dispersion.

2. The Variance and Standard Deviation. The variance of a data set is the average of the squared deviations of each observation from the arithmetic mean. The standard deviation is the square root of the variance. Both of these measures of dispersion are especially useful when one wishes to make inferences beyond the observed data or to obtain a useful measure of sampling precision. For the values 7, 56, 2, 2, 8, the variance is $(7-15)^2 + (56-15)^2 + (2-15)^2 + (2-15)^2 + (8-15)^2/4 = (8)^2 + (41)^2 + (13)^2 + (13)^2 + (7)^2/4 = 2132/4 = 533$. The standard deviation is $\sqrt{533} = 23.1$. (Note: The denominator is usually taken as one less than the number of observations.)

Many other measures of dispersion exist, including mean absolute deviation, semi-interquartile range, etc. For a discussion of measures of dispersion, see Yamane (1964, Chapter 4).

Measures of Association

When more than one quantitative variable is measured, one frequently wishes to describe the type and strength and direction of relationship between those variables. To illustrate, Figure 8-4 presents a plot known as a scatter diagram that illustrates the observed data in Example 4. Note that the relationship between variables appears to be positive and linear. Frequently, product-moment correlation coefficients are utilized as measures of association between quantitative variables. For discussions of correlation coefficients and illustrations of their use, see Walker and Lev (1953, Chapters 10 and 11).

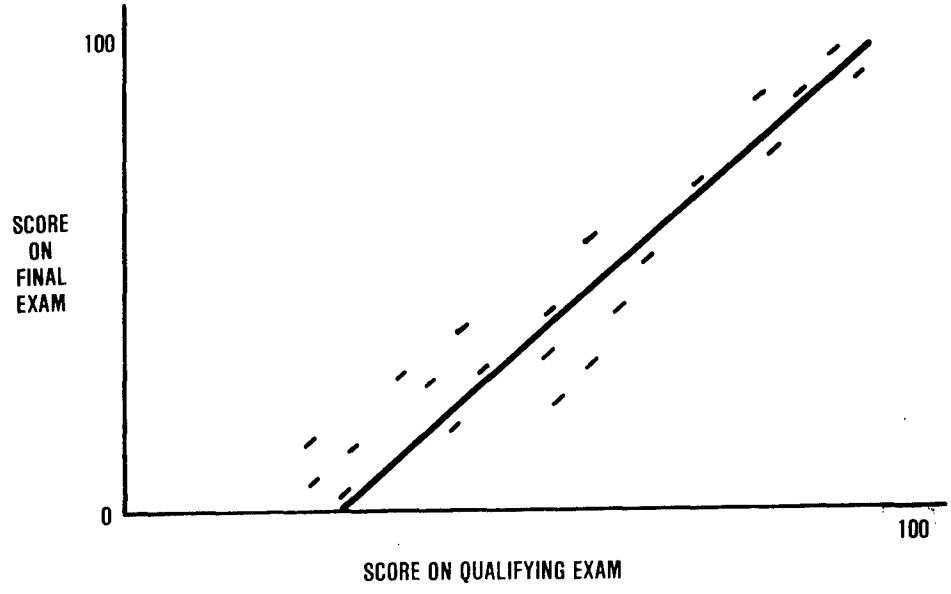


Figure 8-4. Scatter diagram of score on qualifying exam and score on final exam (Example 4).

Other Measures. There exist many other summary statistics for quantitative data. Measures of "skewness" describe the symmetry of a set of observations, while measures of kurtosis consider the "peakedness" of a frequency distribution.

This section was designed merely to introduce the topic of descriptive statistics. It is essential that references such as those cited previously be consulted before attempting to properly summarize any data.

Very often we desire to make statements, estimates, or decisions that go beyond the specific data that have been observed or analyzed. The next section deals with this subject of statistical inference.

Statistical Inference

It is often necessary to generalize findings to a larger domain than the individuals or units actually observed. If we define our population or universe to be that larger domain, and if we define the observations being analyzed to be the sample, we then wish to use the statistics and other information obtained from the sample to make statements about the characteristics of the entire population. This generalization from observed data to population of interest is called statistical inference. Specific areas of statistical inference include:

1. Estimation of Unknown Population Parameters (Note: A parameter is a summary measure of the units of a population whereas a statistic is a summary measure of the units of a sample).

2. Tests of Hypotheses.

3. Analysis of the Relationships Between Variables.

4. Forecasting and Decision Making.

5. Time-Series and Trend Analysis.

The above list is by no means exhaustive; rather, it is intended to provide the reader with some knowledge of the problems with which inferential methodology is concerned. It should be noted that general conclusions derived from sets of observations are necessarily uncertain. Statistical methodology provides techniques for both assessing the accuracy of our estimates and for judging the probability of making incorrect decisions within this climate of uncertainty.

Estimation

Statistical estimation is concerned with the problem of estimating one or more population parameters from the information contained in a sample. One problem that frequently arises in practice is that of estimating the mean of a variable in a population. Another common problem considers estimation of the proportion of individuals or units in a population with some predefined characteristic or membership in a specific category. To illustrate, Example 1 relates to the problem of estimating the average tread life over rough terrain for all tires of a specific type. The technique for analyzing this problem is typical of many statistical analyses; that is, a sample is drawn and summary statistics (obviously including the mean of the sample in this case) are computed from the observations in the sample. The value of the summary statistic (the mean tread life of tires in the sample) is then used as an estimate of the mean tread life of all tires produced. Considering the situation in Example 2, we may wish to estimate the proportion of all Marine recruits (not merely those in the sample) who could use the machine properly after reading the training manual. In this case, the statistic "proportion in the sample who used the machine properly after reading the training manual" is often used as the estimate.

The mere presentation of an estimate, however, avoids the issue of its accuracy or precision or error. Statistical estimation methodology provides techniques for constructing estimates that are, in general, as accurate as can be achieved from the sampling methods employed. Techniques have been developed for estimating the accuracy or error of statistical estimates. These techniques are based upon consideration of the sampling distribution of statistics; that is, the study of the frequency distribution of repeated sampling and estimation from a given population.

Based upon the Central Limit Theorem of statistics, statistical techniques based upon the normal distribution have been developed for judging the precision of estimates of population means and proportions when the sample size is sufficiently large (in most applications, at least 30 to

100) and drawn in a specific manner. Often an estimate and its associated precision are utilized to construct a confidence interval--an interval in which we are reasonably certain that our unknown population parameter is located. For example, rather than state that our estimate of the average tread life of tires is 23,570 miles--we might present our findings as "we are 99 percent 'confident' that the average tread life is between 23,260 and 23,880 miles." For a discussion of these and other estimation concepts, see Dixon and Massey (1969, Chapters 5-7).

Another estimation problem concerns not simply the estimation of the value of an unknown population parameter but, rather, the estimation of differences between a given parameter of two or more populations (i.e., the difference between the average tread life of tires produced by manufacturer A and manufacturer B). For a discussion of this and similar problems, see Dixon and Massey (Chapter 8).

Hypothesis Testing

Hypothesis testing is an aspect of statistical methodology concerned with determining whether an unknown population parameter is equal to a prespecified value (or class of values). Considering Example 2, we might be interested in determining whether the proportion of all Marine recruits who cannot utilize the machine properly after reading the training manual is 5 percent or less. Considering Example 1, we might truly be concerned not with estimating average tread life, but, rather, simply judging whether average tread life is at least 35,000 miles.

Ordinarily, hypothesis testing problems are denoted by specifying both a null hypothesis, or statement, and an alternative hypothesis about an unknown population parameter. Considering our tread life example, our null hypothesis might be "average tread life of all tires is 35,000 miles or more," while the alternative hypothesis might be "average tread life of all tires is less than 35,000 miles." Statistical techniques have been developed for use in deciding which of these hypotheses is correct. These methods enable the user to develop test procedures with definable probabilities of making incorrect decisions.

When one constructs a test of hypothesis, the following two decision errors are possible: (1) Type I Error: Concluding that the alternative hypothesis is true when, in fact, the null hypothesis is true; and (2) Type II Error: Concluding that the null hypothesis is true when, in fact, the alternative hypothesis is true.

The utilization of proper test construction methodologies in conjunction with appropriate sampling techniques allows the user to analyze the probability of making either of these errors. Carrying out a test of hypothesis is called a test of significance. For a discussion of significance testing, see Dixon and Massey (1969, Chapters 6-8).

Multivariate Analysis

Multivariate statistical techniques consider the analysis of several variables at once. This type of statistical analysis is performed when

one wishes to assess the relationship between several variables. Frequently, this type of analysis is utilized when one wishes to search for "causes" (although an analysis of data itself is not sufficient to attribute causality) or for predictive or forecasting purposes. Examples of multivariate techniques include:

1. Correlation Analysis is frequently employed when quantitative variables are analyzed. For example, if our observations consist of pairs of quantitative observations as in Example 4 (i.e., 1st Exam Score, Final Exam Score), correlation analysis may be the appropriate methodology to analyze the relationship between these two variables.

2. Chi-Square Techniques are frequently used to estimate and test hypotheses about the relationships between qualitative or categorical variables.

3. Analysis of Variance (ANOVA) Techniques are often employed when one wishes to analyze the relationship between a quantitative variable and one or more qualitative or categorical variables. An ANOVA might be employed when one wishes to analyze the relationship between test-score (quantitative variable) and region (qualitative variable) and teaching methodology (qualitative variable).

4. Regression Techniques are utilized when one wishes to derive and analyze the relationship between one or more "predictor" or "independent" variable and a dependent variable. For example, one may wish to predict an individual's final exam grade (dependent variable) on the basis of his qualifying exam score (independent variable).

The multivariate techniques discussed thus far are but a handful of the wide variety available. It is critical that no techniques be utilized without a thorough understanding of the assumptions underlying the use of each method. See Snedecor and Cochran (1973); Yamane (1964); Dixon and Massey (1969); or Freund and Williams (1972) for discussion of some multivariate methods.

Time Series

Time series techniques deal with the analysis of the behavior of variables over time. The assessment of trends, cycles, and seasonal fluctuations are some of the questions addressed by time series methodology. See Yamane (1964, Chapters 12 and 13) and Neter and Wasserman (1973, Chapters 29-32) for discussions of time series methodology.

The concepts and methodologies presented in this section are merely an introduction to the type of problems analyzed through statistical inference.

Sampling Techniques

Thus far, we have not discussed techniques for obtaining the data needed for analysis. It is obvious that we wish to obtain and analyze data that are "representative" of the population of interest. Random sampling

techniques have been developed that enable the user to draw samples likely to be representative of the population of interest. These techniques include:

1. Simple Random Sampling. Simple random sampling techniques draw individuals into the sample in a manner whereby every individual or the unit in the population has an equal chance of being selected. Furthermore, individuals are chosen independently. That is, the selection of an individual into the sample has no impact on the selection of any other individual or item into the sample. Therefore, if we wish to select a sample of all Marines using simple random sampling methodology, then we must select individuals so that every marine has an equal (and independent) chance of selection. The utilization of this and other random sampling techniques requires the user to have a frame, or listing, of the individuals or items in the population of interest.

2. Stratified Random Sampling. Stratified random sampling is a sampling technique in which the population is first divided into stratum, or subpopulations, and then a simple random sample is drawn from each stratum. For example, if we divide Marines into stratum based upon geographic location and then sample randomly from each of these stratum, such a scheme would constitute a type of stratified random sampling. Such sampling frequently results in estimates having increased precision. However, stratified sampling is often difficult to carry out in practice.

3. Cluster Random Sampling. Cluster random sampling is a sampling methodology whereby "groups" or "clusters" of individuals or items are selected as part of the sample at once rather than individually. For example, when drawing a sample from the population of all naval personnel serving on ships, we might draw a sample of ships (clusters) and consider all personnel on the chosen ships as members of the sample. Cluster sampling techniques are often the easiest and most inexpensive procedures to carry out. The results of these schemes, however, also tend to be the most difficult to analyze.

Other sampling techniques such as quota sampling, systematic sampling, and combinations of the above schemes are also employed.

It is crucial that data be obtained in accordance with accepted sampling techniques if one wishes to measure the accuracy of estimates, to analyze the probability of making errors, or to make generalizations of findings from sample to population. Furthermore, the specific sampling and estimation (or hypothesis testing) methodology employed enable the user to gauge the size of the sample (number of individuals or units) needed to obtain a desired precision. For discussions of sampling, see McCarthy (1957, Chapter 10) and Kish (1967). For discussions of a related subject, the design of experiments, see Snedecor and Cochran (1973).

Summary

This chapter has introduced some of the basic concepts of statistics--including descriptive techniques, statistical inference, and sampling

methodology. Investigation of references such as those described in this chapter is essential prior to the utilization of any statistical technique.

REFERENCES

Churchman, C., & Ratoosh, P. Measurement-definitions and theories. New York: John Wiley & Sons, Inc. 1959.

Dixon, W. J., & Massey, F. J. Introduction to statistical analysis (3rd ed.). New York: McGraw-Hill, Inc., 1969.

Ellis, B. Basic concepts of measurement. Cambridge, ENG: Cambridge University Press, 1966.

Freund, J., & Williams, F. Elementary business statistics (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc., 1972.

Hamburg, M. Statistical analysis for decision making. New York: Harcourt, Brace, and World, Inc., 1970.

Kish, L. Survey sampling. New York: John Wiley & Sons, Inc., 1967.

McCarthy, P. J. Introduction to statistical reasoning. New York: McGraw-Hill, Inc., 1957.

Neter, J., Wasserman, W., & Whitmore, G. A. Fundamental statistics for business and economics (4th ed.). Boston: Allyn and Bacon, Inc., 1973.

Snedecor, G., & Cochran, W. Statistical methods (6th ed.). Ames, IA: Iowa State University Press, 1973.

Walker, H., & Lev, J. Statistical inference. New York: Henry Holt and Company, Inc., 1953.

Yamane, T. Statistics: An introductory analysis. New York: Harper and Row, Inc., 1964.

SECTION NINE--PERSONNEL PERFORMANCE TEST PLANNER'S CHECKLIST

This section provides a checklist that can be used by test planners to ensure that they have performed all necessary steps to conduct an effective test. The checklist covers necessary questions to be answered for pretest planning, pretest operations, test period, and posttest period. In some cases, comments are provided on the right-hand side of the following pages.

A. PRETEST PLANNING

A1. General Organization of the Personnel Performance Test Plan (PPTP)

a. When must it be completed?

b. Will it be developed in several stages or all at once?

c. Will it be included as one part of the overall Operational System Test (OST) plan or will it be a separate document?

d. How long and detailed must it be?

e. Who must review and approve it?

f. What information must the PPTP contain? (See Section Two)

A2. Equipment/System Background Information

a. Is a description of the equipment/system to be tested available?

(1) Sent along with the test plan requirement?

(2) Are other sources of equipment description specified?

(3) Is the description sufficiently detailed?

b. Is a description of personnel tasks required to operate/maintain the equipment available?

(1) Included with equipment description?

(2) Are other sources of personnel information specified?

(3) Is the description sufficiently detailed?

c. Any other relevant documentation available?

(1) Specifications, e.g., MIL-STD 1472B?

(2) Previous test plans, reports?

A3. Test Purpose

These purposes include:

a. Assess whether system development personnel-related objectives have been accomplished.

b. Ensure that personnel can accomplish required tasks satisfactorily.

c. Forecast how personnel will perform in actual operations.

d. Ensure that personnel elements are effectively integrated into the system.

e. Detect personnel-related deficiencies (e.g., inadequate equipment and procedures, training, technical data, supplies) so that improvements can be made.

f. Identify personnel resources (e.g., manpower) needed to support the operational system.

g. Validate the training program.

Needless to say, not all these purposes may apply to the same OST.

a. Which of these purposes applies to this test? What are the implications of these purposes?

If the purpose is to determine whether system development personnel-related objectives have been accomplished, it is necessary to know what those objectives were--in particular, the criteria for these objectives. The same is true of determining that personnel can do their jobs effectively: how does one know without quantitative criteria? To validate personnel training, again one must know what the training program was designed to accomplish: this means training criteria. To determine what personnel-related deficiencies exist, it is necessary to specify in measuring performance what errors consist of and how inadequate performance is defined.

b. What will be measured by personnel performance tests?

(1) Human engineering.

Have human engineering specifications (MIL-STD 1472B or any other) been levied against the system? Has previous (developmental) testing measured human engineering?

(2) Adequacy of operating procedures.

What does "adequacy" in this context mean? Are procedures up to date? Have they been tried out previously? How do they relate to Marine Corps objectives and operations?

(3) Maintenance operations.

Is maintenance being covered by any other section of the OST test organization? If so, it may not be part of the personnel performance test responsibility unless what the other section is doing fails to involve personnel elements. Is the test concerned with all aspects of maintenance or troubleshooting alone? Has a malfunction reporting procedure already been established? Which of the following maintenance procedures have been established for the system: (a) attempt to repair malfunctions during the test; (b) abort the test and return system to depot; (c) do not attempt to repair but proceed with the test; and (d) working

	around the malfunction. What information should be collected about maintenance (e.g., total down time, repair time, types of errors made)? Examine in detail implications of collecting personnel data in relation to maintenance; they may give the planner a headache.
(4) Correctness of technical manuals.	This aspect is not quite the same as evaluating the adequacy of operating procedures. Determining correctness of any technical manuals involves a great deal of very detailed work.
(5) Adequacy of training.	Training adequacy can be determined in various ways: Types of personnel errors may indicate lack of training; test personnel can be asked directly whether they feel their training was adequate and if not, in what ways it was deficient.
(6) Ability of test personnel to perform tasks.	Comments with regard to performance criteria apply especially to this aspect.
(7) Effect of special operating conditions on personnel performance.	For example, are personnel required to drive tanks at night as well as during the day, in swamps as well as on hard surfaces, etc.? Examine mission objectives to determine if very contrasting operating conditions exist for which it would be useful to collect data on personnel capability to perform under these conditions.
(8) Other (e.g., logistics).	
A4a. Relation of PPTP to System Operations	
(1) What operations will be performed as part of OST?	For various reasons (e.g., cost/time), not all the operations in which the system under test is ordinarily utilized may in fact be tested.
(2) Will personnel data be gathered on all OST operations?	If the system has many operations and personnel, the requirement to collect personnel data on all of these may impose a severe burden on data collectors.

(3) If not, on which ones?

If only selected system operations will be used to collect personnel data, it will be the test planner's responsibility to select these (if not already specified) on the basis of: (a) criticality to mission accomplishment; (b) frequency of performance (more frequently performed operations are, all other things being equal, more important to evaluate); and (c) difficulty of operation, if known (more difficult tasks will stress personnel more).

(4) Are there any special conditions in these operations that would impact on test planning?

(5) Will all tasks performed in specified OST operations be measured?

Each system operation to be measured may require a number of tasks to be performed. Some of these are more or less important, more or less easy to gather data on. The test planner must specify which of these tasks (if not all) must be measured; test observers need this information.

A4b. Measurement of Personnel Performance

(1) What criteria for successful performance of the tasks being measured exist? (List these.)

(2) Are they quantitative and in sufficient detail?

(3) If performance criteria for these tasks are not specified in available documentation, what possible other sources exist?

(4) Can criteria be developed by consensus of experienced personnel? (See Section Two)

A4c. Measurement Methods

(1) Will instrumentation be required? Compare advantages/disadvantages.

In general, unless the data desired can be secured in no other way, instrumentation is not a preferred measurement method because of cost, scheduling problems, the need for specialized equipment operators (and maintainers), and difficulty in using such equipment in a field setting.

(a) If so, is it available?

(b) Must it be procured? From where? What are the procurement procedures?

(c) Cost/schedule.

(2) Observation.

(a) What information will observers record?

(b) What data recording forms will be required?

(3) Interviews of test personnel.

(a) What information will be secured from interviews?

(b) What questions should be asked?

(4) Questionnaires.

(a) What information will be secured from questionnaires?

(b) What questions should be asked?

(5) Ratings.

(a) What information will be collected from ratings?

(b) Who will fill out the rating scales?

(c) Will rating scales have to be developed or are adequate ones available?

A4d. Test Personnel Required

Test personnel are those personnel who operate and maintain the equipment being evaluated.

(1) How many?

(2) When must they be available?

(3) Is special background required? If so, what?

(4) Rank/skill level?

It is not sufficient merely to specify that test personnel will have a given rank and military speciality. Since the personnel will vary in terms of their ability, the planner should ask whether they should come from the top 10 percent in ability, the middle (50%) in aptitude, or even lesser skilled personnel. Obviously, if test personnel are the "cream of the crop" of their speciality, they do not properly represent the great mass of military personnel; however, system performance in the test will be more efficient, since such personnel can more adequately compensate by their skill for any deficiencies the system may have. On the other hand, lower skilled personnel (e.g., the middle 50% in ability) will be more representative of the military population who will eventually have to use the system; but the system in their hands will not look "as good" as if it were operated by more effective personnel. The choice is a matter of philosophy: making the system look its best; or getting results that apply more directly to the overall Marine Corps population.

(5) Secured from what units?

(6) Will test personnel require training on equipment?

(a) If so, will training be given at the factory, by USMC, where, and for how long?

A4e. Test Observers Required.

(1) How many?

(2) What type? Rank/speciality area?

(3) What will their duties be?

It is particularly important to specify in precise detail the activities required of observers. If this is not done, the data recovered may be inaccurate or some may even be missed.

(4) What training will they be required to have? (Describe training.)

It is highly desirable that observers be given realistic training in their observational duties; if this is not done, the data they secure may be inaccurate or some may even be missed.

(5) Who will provide this training, where will it be given, and when?

(6) From where will observers be secured?

(7) To whom will they report?

It is presumed that test observers will report to the individual in charge of personnel performance testing but this should be specified.

A5. Testing Schedule

a. Will personnel performance data be gathered as an integral part of the overall OST?

Ordinarily this is or should be the situation, in which case the overall OST schedule determines the personnel performance test schedule.

b. Will special personnel performance tests be required?

Ordinarily such special-purpose personnel tests should not be necessary if personnel performance testing is fully integrated into all OST phases. However, it is conceivable that special questions relative to personnel may arise that cannot be satisfied in the normal course of OST. The test planner should examine OST operations to be performed before answering this question.

c. If so, what are these and how will they be conducted?

d. What will be the impact of such special tests on the overall test schedule?

A6. Data Analysis

a. What data will be secured?

Before testing begins (even before training of test observers begins), a detailed list of the data items to be collected, along with

information on how and for what purpose they are to be collected should be drawn up so that everyone involved knows exactly what is needed. Statistical analysis cannot be specified before this list is developed.

b. What statistics will be applied?

The plan for statistical analysis of the data is an integral part of the PPTP.

c. How will the data be processed?

d. Who will perform the data analysis?

If the statistical analysis is to be performed by other than the usual USMC agencies, the performing agency should be identified. This includes any personnel who will handle/process the data between the test observer who collects it and the analyzing agency.

e. Are computer facilities necessary? Who will provide them?

A7. Final Test Report

a. What sections will the test report include?

USMC regulations may specify a particular format for the personnel performance test report, whether it is separate or included as part of the overall OST report. Whatever the case, personnel performance test report should have at least the following sections: purpose of test, methods used (including instrumentation (if used), data recording forms, questions asked, etc.), subjects, procedure for collecting data, results, conclusions, recommendations.

b. What is the schedule for the test report?

c. Who will write the test report?

B. PRETEST OPERATIONS

This refers to the period between the time the PPTP is written and the start of actual testing. It includes all the preparations (including observer training) for

conducting personnel performance testing.

B1. Availability of Test Personnel

a. Have they arrived?

b. Have they received required training (including checking out on equipment)?

c. Do they have all required job aids (if job aids are necessary to task performance)?

d. Have they been instructed on the role they will play in OST?

All test personnel should be informed that, as a routine part of OST, their performance will be measured and that they will be interviewed, observed, and/or asked to fill out certain forms. They should be reassured that this evaluation is solely to check out the equipment.

B2. Availablity of Test Observers

a. Are they on-site?

b. Have they received required training as observers?

B3. Availability of Measurement Devices

a. If instrumentation is required, has it been received and checked out and are observers trained in its use?

b. Are all manual recording forms ready and have they been tried out as part of observer training?

c. Are all interview/questionnaire questions developed and tried out in observer training?

d. Is the test schedule up to date?

This will often change up to the start of testing because of delays in getting equipment ready.

C. TEST PERIOD

C1. Initial Checks on First Day's Weeks's) Results

It is highly desirable to check results of the start of testing because various problems often

	arise in testing which must be resolved. If these are not solved, much data collected may be inadequate or even lost.
a. Check with test observers:	
(1) Any difficulties in collecting data experienced by observers or in performing by test personnel?	Observers can usually report on whether test personnel are experiencing difficulties that might interfere with data collection.
(2) Are any changes to measurement procedures, instrumentation, recording forms, or test schedule required?	
(3) What is the effect of such changes on the PPTP?	
(4) Will the desired number of data points be secured?	Results of the first week of testing should indicate whether it is possible to collect all the data specified in the PPTP. Changes, if any, in the overall OST operation (e.g., breakdown of equipment, reshuffling of personnel) may interfere with personnel performance data collection and may require corresponding changes in the data collection procedure.
(5) Are the desired data being secured?	
(6) Do observers appear to know their jobs?	It is desirable to check on how well observers are performing because those who appear to be falling down on the job may have to be reindoctrinated or replaced.
b. Check with test management:	
(1) Is the test on schedule?	
(2) Are any changes anticipated in test operations that will impact on the PPTP?	
C2. Periodic Check During Test Operations Concerning Above Questions	Periodic checks on data collection should be made because at any time the OST operation may be modified (because of equipment malfunction,

scheduling delays) such that personnel performance data collection may have to be curtailed or otherwise modified.

C3. Final Check

a. Have all necessary data been secured and recorded?

If some necessary data are missing at the conclusion of OST, the personnel performance test planner will have to decide what can be done about this.

b. Have all data been transmitted to data analysts?

D. POSTTEST PERIOD

D1. Data Analysis

a. What is the schedule for data analysis? Is the analysis on schedule?

b. Are the data appropriate to the planned statistical analysis?

If, for various reasons, it is found that the data collected will not fit the planned statistical analysis, important decisions about changing the analysis format must be made.

c. Were sufficient data collected to satisfy test objectives?

d. Are the results relevant to the test objectives?

Inadequate planning may result in insufficient or irrelevant data being collected. If so, critical decisions must be made.

D2. Final Test Report

a. What is the schedule for the preliminary draft? For the final test report? Is the report writing on schedule?

b. Are the results/conclusions clearcut?

c. What recommenations can be made? Are they reasonable? What will their impact on the system be?

d. What system modifications are required:

(1) In hardware?
(2) In procedures?
(3) In training?
(4) In manning?

SECTION TEN--THE PERSONNEL PERFORMANCE TEST REPORT

This section outlines the major points to be included in a personnel performance test report.

Introduction

This section describes the major points to be included in a report describing the results of the personnel performance test. The Marine Corps has its own report format, as described in MCO 5000.11, Test and Evaluation of Systems and Equipment for Operating Forces of the Marine Corps; and the personnel performance test report described in this Section should conform to that directive. Within the constraints of the Marine Corps test report format, the items described herein should be included.

The personnel performance test report is a major vehicle for the transmission of information about the test and will reach a wide variety of interested agencies. It is important, therefore, that care be taken in its preparation.

The major categories which the test report should cover are:

1. Summary
2. Test Objectives
3. Test Method
4. Results
5. Conclusions
6. Recommendations
7. References
8. Appendices

Outline of the Personnel Performance Test Report

A. Summary of Test Report

A paragraph or two describing the highlights of the study with emphasis on:

1. Purpose of the personnel performance measurement.
2. When and where test was conducted.
3. Major results and conclusions.

B. Personnel Performance Test Objectives

1. This section should describe the objectives for which the personnel performance test was conducted. Specifically these objectives should have been to answer the following questions:

a. The determination of how well personnel perform with the new system.

b. The determination of whether personnel satisfy system requirements as far as their performance is concerned.

c. The problems that personnel experience as these reflect on various aspects of the system, e.g.:

(1) Human engineering of equipment design.
(2) Operating/maintenance procedures.
(3) Manning.
(4) Appropriate personnel background to perform duties.
(5) Training.
(6) Other (e.g., logistics, manuals, job aids).

C. Test Method

1. Test Personnel

This section descibes the characteristics of personnel acting as test subjects.

a. Definition of test personnel as those operating and maintaining the system during test exercises.

b. Selection of test personnel:

(1) Personnel selected from what units.
(2) Personnel background (e.g., military speciality, rank).

Indicate any personnel characteristics particularly important to the system (e.g., strength, aptitude).

(3) Number of subjects.
(4) Selection criteria.

Indicate the basis for determining how many personnel were selected as test subjects and the rationale for the selection criteria (e.g., the 95th percentile of scores in school training, ranking by commanding officer of their unit, selection on a random basis). If personnel were selected by tests or scores, what were these? Were there any constraints on personnel selection (e.g., small population) and what were these?

c. Special training received by test personnel (to operate/maintain test system).

2. Test Procedure

This section describes general test methodology and performance criteria.

a. Test was conducted over what time period? Using what facilities? As part of operational exercises or in the form of special tests? How was test conducted?

b. Tasks/operations for which personnel performance data were collected. List and, if reader is unlikely to be familiar with these, describe major functions/tasks performed for which data were collected. If not all tasks/operations were observed/measured, what was basis for selection? Indicate number of operating cycles (e.g., tank runs, rounds fired) on which data were collected.

c. Experimental Design

If a specific experimental design was used (e.g., repeated measures on the same subjects, special order of performing tasks such as alternating day/night exercises), describe at this point and indicate rationale for the design.

d. Specific variables tested (e.g., day vs. night operations, sandy vs. marshy terrain). Reason for being concerned about these variables.

e. Personnel performance criteria:

(1) For all major operational tasks performed, what quantitative criteria describe adequate personnel performance (e.g., allowable firing miss distance (2 feet); maximum time allowed for replacing X component (38 minutes))?

(2) Indicate source of criteria:

(a) Overall system requirements.
(b) System documentation (reference).
(c) Operational requirements determined by mission.
(d) Consensus of skilled experts.

(3) List any objective performance measures collected and categorize these by the criteria in section C2e(1). Define each measure employed (e.g., what is meant by error, response time, etc.?).

(4) Indicate any difficulties or problems in measuring these criteria. If so, what was done to resolve these problems?

f. Data collection methods:

(1) If observation was used, indicate:

(a) Who made the observations.
(b) How the observations were made.
(c) How observers were qualified to make these (e.g., training, experience).
(d) What observers were supposed to observe in relation to what system operations.
(e) If any observational data recording forms were used, place these in the appendix.

(2) If interviews were held with test personnel, describe:

(a) The general content of the interviews.
(b) When and where held.
(c) Who was interviewed (not in terms of specific names but in terms of categories of personnel).
(d) Average length of interview.
(e) Whether taped or manually recorded.

(3) If questionnaires were used, include the form used in appendix, and describe:

(a) The general content of the questionnaire.
(b) As with interviews, when or how frequently the questionnaire was employed.
(c) Who completed questionnaires.

(4) If rating scales were used, include them in appendix, and describe:

(a) The nature of the scales.
(b) The data they were supposed to produce.
(c) Who completed scales and how frequently.

(5) If instrumentation was used to collect objective measures, describe:

(a) The general nature of the instruments (e.g., time and events recorder, noise level measurement device).
(b) The particular measures it was used for.

Note. If the instrument is novel, it might be advisable to append a more detailed specification of its operating characteristics, including a photograph.

(6) If the experimental design of the study (see section C2c) involved a comparison of two or more conditions (e.g., performance under different climatic or terrain conditions), include a description of these special conditions. Any special conditions that were important to the test should be described in detail.

D. Results

1. Statistical Analysis

Referring back to the experimental design (section C2c) as the rationale for the analysis:

a. Describe the analyses performed (e.g., Analysis of Variance, t-tests, correlations).

b. Indicate the adequacy of the data collected, particularly any factors that might have affected the analysis, such as too little data, non-normal distribution, etc.

Note. If the manner in which the overall OST was performed influenced the quantity/quality of the data, indicate what this was.

2. Personnel Performance Effectiveness

a. Objective Measures.

This section refers back to section B1a and describes how well personnel have performed. It should include data gathered by instrumentation

or by observation of quantitative indices (e.g., miss distance in firing at targets). It should include both operator and maintainer functions unless maintenance will be covered in a separate report or report section.

(1) Determine the statistical mean (average) and standard deviation performance in terms of specific measures for each major function/task as previously called out in section C2b. Compare the mean with any system-required personnel performance (the standard of accomplishment). Examine the variability (standard deviation) of the performance: Is the variability so great that the mean value is unreliable?

(2) Determine the statistical significance of differences between any conditions being compared (section C2d).

(3) List the performance values for each major function/task in tabular form. If these data are extensive, they should be included in a separate appendix.

(4) Where appropriate, categorize types of errors made by personnel and indicate their frequency.

Note: The statistical section of the report should be written by a qualified statistician or at least reviewed by him.

b. Subjective Data

Any subjective data (i.e., those gathered from observations, interviews, questionnaires, ratings, or critical incidents) that bear on how well personnel have performed or which explain their performance should be included here. Subjective data which can be described in quantitative terms (e.g., mean and standard deviation of ratings, the percentage of those responding yes and no to particular questions in interviews and questionnaires, the number of those observed to perform in particular ways or the frequency of their performance) should be listed in tabular form, where possible.

3. Equipment Characteristics

This section describes any human engineering equipment discrepancies that have been noted by test observers or by test personnel in interviews, questionnaires, or rating scales.

a. List each discrepancy per equipment and refer to appropriate section of MIL-STD 1472B (Department of Defense, 1974) for which it is a discrepancy. For example, "the noise level within the tank compartment is excessive, measuring peaks of 90db (see paragraph 5.8.3.2 of MIL-STD 1472B)."

b. Indicate the importance of the discrepancy in terms of its effect on test personnel and/or mission accomplishment, using a scale such as (1) minor--1, (2) moderately important--2, (3) extremely important--3. Indicate actual or possible effects on performance from the test data.

c. Where appropriate, include diagrams, photos, etc., illustrating the discrepancy (e.g., diagram of improperly laid out control panel).

4. Operating and/or Maintenance Procedures and/or Manuals

List any inaccuracies or changes required in procedures or manuals that were found as a result of test performance.

5. Training

The purposes of this section are to describe the adequacy of the training given test personnel to operate/maintain the new system and to indicate where further training is required. The training curriculum provided test personnel (see section C1c) should be examined in terms of how well personnel performed and how they felt about their training. Data will be derived from a number of sources:

a. Functions/tasks with inordinately high error rates or very delayed response times, where the cause of such errors appear to result from inadequate training.

b. Data secured from interviews and questionnaires in which questions were asked specifically about training (e.g., were there any functions/tasks for which not enough training was given or the training appeared to be inappropriate?).

6. Personnel Requirements

This section includes any deficiencies noted in:

a. Manning--the number of men required to operate/maintain the system (for example, if two men are specified but three are required or vice versa).

b. Special aptitudes noted that are required to perform system functions.

E. Conclusions

1. General

This section describes the answers to objectives in section B1. System personnel can or cannot operate/maintain the system to requirements. Manning is or is not appropriate for required tasks. Training is or is not adequate, etc.

2. Specific

Inadequacies were found in:

a. The following tasks (list and describe).
b. Human engineering (describe).

c. Personnel requirements (describe).
d. Training (describe).

These have the following effects on system operations (describe).

F. Recommendations

1. Changes to the system should be made with regard to:

a. Equipment design.
b. Procedures.
c. Personnel requirements.
d. Training.
e. Other.

2. Indicate which of the above modifications can be made by:

a. Equipment redesign.
b. Changes to procedures.
c. Training of personnel.
d. Logistics (e.g., spares, tools, etc.)

G. References

1. Military documents cited.
2. Civilian publications cited.

H. Appendices

1. Tabular data (e.g., statistical analyses, lists of errors made, important raw data).

2. Photos/diagrams of important items of equipment referred to in the body of the report.

3. Data collection forms, interview questions, etc.

SECTION ELEVEN--USEFUL REFERENCES

This section presents additional specifications and reference materials which the evaluator may find useful.

Specifications and Standards

MIL-H-46855A Human Engineering Requirements for Military Systems, Equipment and Facilities, 2 May 1972.

MIL-STD-1472B Human Engineering Design Criteria for Military Systems, Equipment and Facilities, 31 December 1974

MIL-STD-721B Definitions of Effectiveness Terms for Reliability, Maintainability, Human Factors, and Safety, 25 August 1966.

Reference Books

McCormick, E. J. Human factors engineering (3rd edition). New York: McGraw-Hill, 1970.

Meister, D., & Rabideau, G. F. Human factors evaluation in system development. New York: Wiley, 1965.

Van Cott, H. P., & Kinkade, R. G. Human engineering guide to equipment design. Washington, D. C.: U. S. Government Printing Office, 1972.

Woodson, W. E., & Conover, D. W. Human engineering guide for equipment designers (2nd edition). Berkeley, CA: University of California Press, 1966.

SECTION TWELVE--INDEX

DISTRIBUTION LIST

Chief of Naval Operations (OP-987H), (OP-964D)
Chief of Naval Personnel (Pers-10c), (Pers-2B)
Chief of Naval Material (NMAT 04), (NMAT 08), (NMAT 08T244)
Director of Navy Laboratories
Commandant of the Marine Corps (5)
Commanding Officer, Naval Development and Training Center (Code 0120)
Director, Training Analysis and Evaluation Group (TAEG)
Army Research Institute for the Behavioral and Social Sciences
Military Assistant for Training and Personnel Technology, Office of the Under Secretary of Defense for Research and Engineering
Coast Guard Headquarters (G-P-1/62)