AIAA
G-035A-2000

Guide

Guide to Human Performance Measurements

AIAA
G-035A-2000

Guide to Human Performance Measurements

Sponsored by

American Institute of Aeronautics and Astronautics

Abstract

This Guide provides methods for measuring human performance for the purpose of scientific research and system evaluation. The guidelines are intended to assist scientists and systems specialists in selecting human performance measurement methods appropriate to the situation being studied or the system being evaluated.

Library of Congress Cataloging-in-Publication Data
Guide to human performance measurements / sponsor, American Institute of Aeronautics and Astronautics
p. cm.
"AIAA G-035A-2000"
ISBN 1-56347-451-4 (softcover)
1. Human engineering--United States--Measurement. 2. Performance standards--United States. I. American Institute of Aeronautics and Astronautics
TA166 .G85 2000
620.8'2'0287--dc21 00-040175
CIP

Published by
American Institute of Aeronautics and Astronautics
1801 Alexander Bell Drive, Reston, VA 22091

Printed in the United States of America

Contents

List of Figures

List of Tables

Foreword

The project on human performance measurement was initiated under the auspices of the Life Sciences and Systems Committee on Standards (LS&S / CoS) of the American Institute of Aeronautics and Astronautics (AIAA).

A standard treating human performance measurement is complicated by the area of application, the level of detail required, the type of testbed available, and a host of performance-shaping functions. The objective of this standard is to guide the reader regarding these and other human performance issues as well as to suggest measurement methods and metrics.

Development of the *Guide to Human Performance Measurements* began in response to an AIAA call for standards proposals. Dr. Valerie Gawron was appointed chair of the LS&S / CoS, and the first proposal was approved. Dr. Gawron then recruited experts in the field of human performance measurement. These experts assembled in January 1990 to prepare an outline for this standard. The outline was expanded and reviewed at a second meeting in March 1990. The LS&S / CoS approved the final draft of the first edition in May 1991.

The first revision was undertaken in June 1999. At the request of the CoS Chair, it was proposed to separate Appendix B from the rest of the standard in the interest of obtaining broader circulation. AIAA, as the source of that material, is recognized in the new publication, *Human Performance Measures Handbook*, Lawrence Erlbaum Associates, Mahwah, NJ. The pair of documents can be obtained from AIAA. The references were updated and expanded. It is expected that this standard will continue to be refined based on user comments and improvements in human performance measurement techniques.

The following organizations, recognized as having an interest in human performance measurement, were contacted during the development of this standard. Inclusion in this list does not imply that the organization approved the document formally.

Government Organizations

Air Force Flight Test Center
Air Force Operational Test and Evaluation Center
Air Force Research Laboratory
Army Human Engineering Laboratory
Army Research Institute
Department of Defense Human Factors Engineering Technical Group
Federal Aviation Administration
National Aeronautics and Space Administration
Naval Air Warfare Center

Professional Organizations

Aerospace Medical Association
American Psychological Association
Association of Aviation Psychologists
Human Factors and Ergonomics Society
Military Operations Research Society
Society of Flight Test Engineers

The AIAA Standards Procedures provide that all approved Standards, Recommended Practices, and Guides are advisory only. Their use by anyone engaged in industry or trade is entirely voluntary. There is no agreement to adhere to any AIAA standards publication and no commitment to conform to or be guided by any standards report. In formulating, revising, and approving standards publications, the Committees on Standards will not consider patents which may apply to the subject matter. Prospective

users of the publications are responsible for protecting themselves against liability for infringement of patents, copyrights, or both.

When this standard was revised, the AIAA Life Sciences and Systems Committee on Standards included the following members:

Valerie J. Gawron, Ph.D., Chair (Veridian)
Frank Bick (FAA)
Leonard Cipriano (Lockheed Martin, retired)
Richard E. Cordes, Ph.D. (IBM)
June Ellison (NASA Headquarters)
Edwin Fleishman (George Mason University)
Ed Lehman (Veridian)
James C. Miller, Ph.D. (Air Force Research Laboratory)
Deborah Moisio (Air Force Flight Test Center)
John M. Reising, Ph.D. Air Force Research Laboratory)
Samuel Schiflett, Ph.D. (Air Force Research Laboratory)
Jennifer Snodgrass (Tinker AFB)
Vernon E. Strength, Ph.D. (Boeing Space & Communications)
Ching Tsai (Boeing Space Systems)
Donald Vreuls (VRS Corporation)

This document was approved by the Life Sciences and Systems Committee on Standards in February 2001.

This document was accepted for publication by the AIAA Standards Executive Council in January 2001.

1 Scope and Purpose

1.1 Scope

This Guide suggests methods for measuring human performance for scientific research and system test and evaluation. The information contained in this document is provided as guidance, not mandated as direction. This Guide should be considered during the planning, conduct, and analysis of human performance measurement activities. The objectives of this Guide are: (1) to foster human performance measurement (HPM) techniques that have proved to be effective; (2) to promote commonality across research projects and, thus, enable comparison of results across evaluations; and (3) to develop and use common HPM tools for data collection and data processing.

Note that this Guide does not include physiological measures. Such measures will be handled in a separate AIAA standard. Until then, the reader is referred to Martin and Venables (1980).

1.2 Purpose

The purpose of this Guide is to aid the reader in measuring human performance. The reader is assumed to have a basic knowledge of experimental design, statistics, and human performance.

2 Vocabulary

anemometer An instrument used to measure local air flow.

anthropometer An instrument consisting of a ruler with two perpendicular movable legs, used to measure distances such as upper-arm length.

asymptotic learning The point at which performance does not improve with increased practice.

average dwell time A measure of instrument-scanning behavior: “the total time spent looking at an instrument divided by the total number of individual dwells on that instrument” (Harris, Glover, and Spady, 1986, p. 38).

consensus A substantial agreement reached by directly and materially affected interests.

continuous performance Performance, such as tracking or monitoring, that requires constant attention over a period of time.

CRT Cathode-ray tube.

DAT Digital Audio Tape.

dimensions Units of measurement, e.g., deviation from glideslope in meters.

discrete performance Performance that has a well-defined start and end, such as switch activation or issuance of a voice command.

DT&E Developmental test & evaluation

dwell percentage A measure of instrument-scanning behavior: “dwell time on a particular instrument as a percent of total scanning time” (Harris, Glover, and Spady, 1986, p. 38).

dwell time A measure of instrument-scanning behavior: “the time spent looking within the boundary of an instrument” (Harris, Glover, and Spady, 1986, p. 38).

element The smallest logically definable unit of behavior required for completing a task or step, e.g., verify that rpm is between 4500 and 6000 (Berson and Crooks, 1976).

fixation A measure of instrument-scanning behavior; a series of lookpoints that stay within a visual radius of 1 degree (Harris, Glover, and Spady, 1986, p. 38).

fixations per dwell A measure of instrument-scanning behavior: "the number of individual fixations during an instrument dwell" (Harris, Glover, and Spady, 1986, p. 38).

function A major category of activity associated with a system or subsystem and assigned to a person or a machine or shared between a person and a machine (Berson and Crooks, 1976).

gas tester An instrument that measures, detects, and quantifies presence of specific gases.

goniometer An instrument for measuring angles, including human joint angles such as wrist flexion.

HPM Human performance measurement.

hygrometer An instrument that measures relative humidity.

LED Light-emitting diode.

lookpoint A measure of instrument-scanning behavior, e.g., "the current coordinates of where the pilot is looking during any one thirtieth of a second" (Harris, Glover, and Spady, 1986, p. 38).

oculometer An instrument that measures lookpoint.

OJT On-the-job training.

one-way transition A measure of instrument-scanning behavior: "the sum of all transitions from one instrument to another (one direction only) in a specific pair" (Harris, Glover, and Spady, 1986, p. 38).

OST Operational system testing

OT&E Operational test & evaluation

out of track "A state in which the oculometer cannot determine where the pilot is looking, such as during a blink or when the subject's head movement has exceeded the tracking capabilities of the oculometer" (Harris, Glover, and Spady, 1986, p. 38).

photometer An instrument for measuring luminous intensity, luminous flux, illumination, or brightness.

protractor An instrument for measuring angles.

PSF Performance-shaping factor.

psychrometer A hygrometer consisting essentially of two thermometers that measure the dryness of the atmosphere.

saccade A measure of instrument-scanning behavior: "the spatial change in fixations" (Harris, Glover, and Spady, 1986, p. 38).

scan "Eye movement technique used to accomplish a given task. Measures used to quantify a scan include (but are not limited to) transitions, dwell percentages, and average dwell times" (Harris, Glover, and Spady, 1986, p. 38).

sliding caliper An instrument consisting of two curved, hinged legs, used to measure external dimensions.

sound analyzer An instrument that provides octave-band analysis.

sound pressure level meter An instrument that measures steady-state sound.

spot brightness meter An instrument that measures small-area brightness.

spreading caliper An instrument consisting of two curved, hinged legs, used to measure internal dimensions.

step Activities (perceptions, decisions, and responses) that fulfill a portion of the immediate purpose within a task. Alternatively called a subtask (Berson and Crooks, 1976).

task The composite of related activities (perceptions, decisions, and responses) performed for an immediate purpose, e.g., take-off from an airfield (Berson and Crooks, 1976).

thermometer An instrument that measures air, surface, or liquid temperature.

transition A measure of instrument-scanning behavior: "the change of a dwell from one instrument to another" (Harris, Glover, and Spady, 1986, p. 38).

transition rate A measure of instrument-scanning behavior: "the number of transitions per second" (Harris, Glover, and Spady, 1986, p. 38).

two-way transition A measure of instrument-scanning behavior: "the sum of all transitions between an instrument pair, regardless of direction of the transition" (Harris, Glover, and Spady, 1986, p. 38).

vibration analyzer An instrument that analyzes amplitudes of vibrations at selected frequencies.

vibration meter An instrument that measures amplitude and frequency components of complex vibrations.

workload The effort expended by the human operator in accomplishing the imposed task requirements.

3 Applications

3.1 General Applications

This Guide is appropriate for a variety of behavioral specialties, technologies, and applications.

The specialties include: (1) human-factors engineering, in which measurements are made to determine whether the physical configuration of equipment or a system is optimum for human control and operation; (2) training, in which measurements are made to determine the effects of instruction on personnel performance or to determine the instructional variables affecting performance; and (3) test and evaluation, in which measurements are made to evaluate the capability of the human to operate and maintain the system in its intended environment.

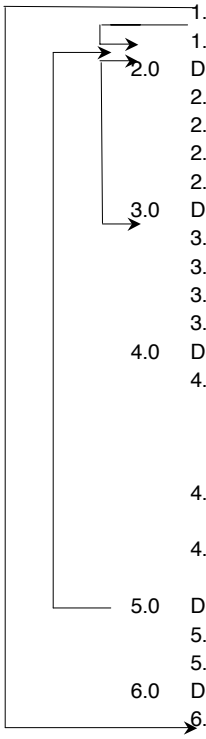
The measurements described in this Guide can be applied to any technology that is controlled and operated by personnel (system personnel) or that affects humans as clients, even if these are not system personnel, or both. For example, the operators of a computerized production line stamping out automobile bodies are system personnel; the owner of one of these automobiles is a client of that system. The focus of the measurements is on the human's performance, although it may appropriate to also measure the equipment or system with which he or she interacts.

These measurement guidelines can be applied to a wide range of devices, from the individual work station (such as a word processor) to a large system (such as an airliner). They can be applied to the single operator and to teams of varying size. More specific applications are to types of measurement: (1) basic and applied research; (2) system design and development; and (3) system test and evaluation. Discussions of these applications are provided below.

All measurements should include consideration of the system context in which they are made. System context is defined as the next higher level system in which the equipment or system being tested is embedded. For example, if measurements are made for a word processor, the office in which the word processor is located should be examined before testing to determine if factors present in the larger office could influence operator performance with the word processor.

Because this Guide can be used in any situation in which humans control or are influenced by equipment and systems, it is independent of any particular application. It is, therefore, incumbent on those who use this Guide to ensure that their usage in any specific context is both scientifically and legally appropriate. Measurements involving human subjects are controlled by ethical standards described in such documents as the Code of Federal Regulations, Part 46, which forbids exposing human subjects to harmful or potentially harmful tests. These documents also require that all research involving human subjects must meet current ethical standards and be approved by authorized review boards for the protection of human subjects.

Before proceeding to a detailed description of various measurement aspects, it will be useful to summarize the major steps in measurement (Figure 1). Note in Figure 1 that there are feed-forward and feed-back loops and, in particular, that it may be necessary to revise measurement objectives and procedures when required resources are not available.



1.0 Define the system to be measured in terms of
- 1.1 Specific equipment
- 1.2 Specific tasks
- 1.3 Questions to be answered (see Section 4.3 and Sections 3.2, 3.3, 3.4, and 3.5)

2.0 Determine measurement objectives
- 2.1 Basic research (see Section 3.2)
- 2.2 Applied research (see Section 3.3)
- 2.3 System design and development (see Section 3.4)
- 2.4 System test and evaluation (see Section 3.5)

3.0 Determine the measurement environment (see Section 5.1)
- 3.1 Laboratory
- 3.2 Engineering facility
- 3.3 Field test site
- 3.4 Operational environment

4.0 Determine resources required
- 4.1 Subjects
  - 4.1.1 Number
  - 4.1.2 Type
  - 4.1.3 Availability; if not available, reconsider 1.0 and 2.0
- 4.2 Cost
  - 4.2.1 Money available? If not, reconsider 1.0 and 2.0
- 4.3 Time required for measurement
  - 4.3.1 Available or not? If not, reconsider 1.0 and 2.0

5.0 Determine measurement constraints
- 5.1 None of importance
- 5.2 Important; re-examine 2.0

6.0 Determine criteria and measures
- 6.1 Criteria (see Sections 4.2.2 and 4.4)
- 6.2 Measures (see Section 5.2)
  - 6.2.1 Validate by expert opinion and/or pretest and comparison of pretest data with objectives
  - 6.2.2 Sufficiently sensitive/objective?

7.0 Develop measurement procedures
- 7.1 Examine available procedures (e.g., instrumentation, observation, interviews) (see Sections 3.4.2, 3.4.3, 3.4.4, and 4.7)
- 7.2 Select procedures in terms of all preceding steps

8.0 Develop statistical analysis plan (see Section 4.8.5.3)
- 8.1 If measurement procedures can provide statistically valid data, accept procedures
- 8.2 If measurement procedures will not provide statistically valid data, revise procedures

9.0 Document all procedures by writing test plan (see Section 3.5)

10.0 Pretest total measurement procedures (see Section 3.5)
- 10.1 Accept unaltered
- 10.2 Unacceptable; revise procedure

11.0 Collect data (see Section 4.8)
- 11.1 Apply data to 1.3 and 2.0
  - 11.1.1 If questions and objectives answered, stop measurement
  - 11.1.2 If questions and objectives unanswered, continue measurement

12.0 Analyze data (see Section 4.8.3)

13.0 Write test report

Figure 1 – Major Steps in Measurement

3.2 Basic Research

Basic behavioral research is performed to develop an understanding of how behavioral variables function, the interrelationships among them, and their effect on the performance of individuals and groups; and, through the latter, their effect on systems.

As a means of acquiring knowledge, basic research is often theory-directed, performed for either the development of a theory or testing the validity of a theory. Basic research can also be initiated simply to fill a gap in knowledge; the researcher says, in effect, we don't know enough about the affect of this factor on task performance.

Basic research melds into applied research. If the research goal is simply to understand, the research is basic. If the goal is to make some specified use of the information gained by the research, it is applied. In any particular case, however, the distinction may be very fine.

The type of questions that the researcher asks (see also 4.3) determines the type of research performed. Questions appropriate for basic research include:

(1) What is the nature of human performance? When one merely asks "what is going on?", the measurement produced is merely descriptive. For example, if one simply wishes to determine the speed with which a skilled carpenter can hammer three-inch nails into a 2x4 plank, this is purely descriptive, with no attempt to consider factors that may influence this speed. Some basic research is of this type, but most basic research involves relationships, as can be seen from the following question.

(2) How does variable X affect task performance? If the researcher were concerned about how ambient lighting affected speed of nail driving, he or she would ask question 2. In the process of answering question 2, question 1 would also be answered, because nail-driving speed would need to be determined in any case, since speed is the dependent variable or measure in both questions (see also 4.6).

(3) What is the relationship between variable X and variable Y? This is an expansion of question 2, since two independent variables (see 4.6) would be involved. One might, for example, be interested in the relationship between ambient lighting and carpenter experience and their individual and combined effects on nail-driving speed. In asking these more complex questions, the researcher usually has a theory about what would happen and selects his or her variables to reflect that theory. For example, the researcher might consider that inexperienced carpenters would need more light than experienced carpenters, because the latters' greater experience might compensate for lesser amounts of light.

(4) What is the effect of human task performance on the larger system performance in whose context the former takes place? As pointed out previously, one can look at the individual performance one is measuring alone or one can examine that performance to see if it had an effect on the larger system. Carpentry is only one part of many functions performed in erecting a house, for example. If, in answering question 3, one examined the effect of experience on nail driving, one might go one step further and ask whether rough carpentry experience had some effect on the speed with which a three-bedroom house was erected. It would be necessary to contrast carpenters of varied degrees of experience working on the same type of three-bedroom house. Such a contrast is usually called an experiment (see 4.6) and requires specific manipulations of contrasting conditions. Although it is theoretically possible to find, in the real world, carpenters with differing amounts of experience working on the same type of house, such natural experiments are unlikely.

Since basic research usually examines the effects of individual and combined variables, it becomes necessary to extract these from the overall measurement situation. So, the amount of ambient lighting would need to be varied carefully, and carpenters with different amounts of experience would need to be assigned to different degrees of lighting. This can be done only by the researcher exercising control over the lighting and the work experience of his or her subjects. The greatest degree of control can be found in the laboratory, which is why most basic research is performed there. However, there are situations in which the measurement cannot be performed in a laboratory, and other testbeds will be necessary.

Moreover, descriptive measurement can often be performed in the operational environment. Since, in descriptive measurement, one is not examining variables, it is often possible to find and measure the phenomenon of interest in the real, operational world. Thus, to determine nail-driving speed without considering the factors that affect it, it would only be necessary to find carpenters working on a house and measure their speed with a stop watch.

3.3 Applied Research

As indicated in the previous subsection, applied research (or, as some call it, problem-solving research) has a more specific goal beyond simple understanding of the variables affecting human performance. The most effective applied research should also consider in its design the relevant variables, so that the conclusions reached would be more generalizable. Applied research then consists of investigation of the behavioral factors affecting or related to the use of technology, technological influences on human performance, or both.

Although applied studies have a more or less specific goal, the research need not be performed on a specific, already existent system or a system development project. The latter is considered in the next subsection under developmental testing. An already existent system may be used as the framework of the applied research, but the system does not drive the research—the objective does. For example, much applied research in World War II was performed on cathode-ray tubes (CRTs) already developed for a specific radar system, but the intent was to determine the effect of various technological variables on detection and tracking capability. The research was to be of value to all classes of a particular technology, to human performance in general, or to both.

For example, during World War II, government engineers conducted research on the relative visibility of radar “blips” to develop a screen design that would be relatively error-free. They also studied various knob shapes to determine how discernible each was, so that pilots would be less likely to confuse one control with another. More recently, studies have been performed to determine the utility of color-coding CRTs, or the relative effectiveness of varying teaching methods. The results of all these could be applied in different contexts or in different systems.

The kinds of questions one asks in applied research are:

(1) What is the effect of a technological variable on human performance? For example, can synthesized speech in computers be clearly understood by operators? The application is to the development of more adequate synthetic speech.

(2) What is the effect of a behavioral variable on the use of technology? For example, what minimum CRT resolution is necessary for an operator to read a TV signal? The application is to the design of better TV sets.

(3) As for two or more behavioral methods, techniques, or processes, which is most effective for achieving a technologically related goal? For example, which is more effective in securing knowledge from a human expert: a personal interview or an automatic (computer-driven) interview? Clearly, the results of the study will apply to the entire knowledge engineering process.

(4) Given two or more technologically related methods, techniques, processes or entities, which is more effective in achieving desired human performance? For example, are aircraft simulators as effective on the whole as an equal length of time spent in actual flying? If a sufficient number of studies addressing this question were performed, the results would be significant for the development of aircraft simulators.

The application of technology and technology-related factors helps to differentiate applied from basic research, but the distinction between the two is less important than knowing what questions to ask.

Applied research, like basic research, also requires control, although this control may be less critical. In general, however, applied research is also performed in a laboratory, simulator, or field test facility, rather than in the operational environment.

3.4 System Design and Development

3.4.1 Scope of this Application

Measurement in the form of testing and evaluation is an integral part of system (or equipment) design and development. It runs throughout development and operation. The continuous testing occurs because design is iterative and recursive and occurs in increments. Testing is necessary, because design is the solution of an engineering problem; so, the designer must determine whether the solution is adequate to the requirements imposed. For a discussion of testing and evaluation in the developmental context, see Meister (1985, 1986).

A distinction must be made between conceptual and empirical testing. If an engineer examines a design drawing or a computer state diagram to ascertain its adequacy from a human performance standpoint, using learned design guidelines, he or she is performing a conceptual test. The difference between this type of testing and empirical testing is that, in conceptual testing, the human performance in the context of which the test is conducted is anticipated or hypothesized human performance; in the empirical test, there are actual human subjects who perform. Most conceptual testing is very private, although there are certain techniques, such as the walkthrough, that are usually performed publicly. Conceptual test techniques include checklist evaluation and the walkthrough; empirical test techniques include mockup testing (hardware systems), rapid prototyping (software systems), and full-scale mission performance (simulators, operational system testing (OST), and experimental testing). These techniques will be described later.

The kinds of measurement questions that are addressed in system development are:

(1) For two or more design configurations, which is best from a behavioral performance standpoint?

(2) Does an equipment or software system configuration conform to accepted behavioral design standards? These standards may be either formal and written, as in Mil-Std-1472, or may be implicit in the evaluator's personal knowledge base.

(3) Does the equipment or software produce acceptable human performance? This question, which is also the primary question asked in system test and evaluation, implies the existence of a behavioral performance standard for the system considered, e.g., a maximum time to perform, a maximum number of errors made by personnel, an accuracy or quality requirement, or all of these combined. In most cases, this standard is not explicit but is part of the developer's or customer's knowledge base.

System design and development proceeds through a series of phases that meld into each other so that only phases widely separated in time can be clearly distinguished. Moreover, design is both iterative and recursive, meaning that sometimes test results will force the designer to return to an earlier design phase to modify his or her design. Both hardware and software development have parallel phases, although distinctly different activities are performed, e.g., mockup testing in hardware, rapid prototyping in software. In both hardware and software, the need for the new system and the constraints development will face (e.g., time, money, or technological capability) are examined. Alternative general design configurations are considered and one selected. All of this takes place in what can be called *predesign.* Development then enters a phase of intensive, progressively more detailed design, in which testing assumes a critical role. The system enters the production phase when the first operational prototype is tested (see 3.5).

Developmental testing in an engineering context, whether hardware or software, is often constrained by exigencies of time and lack of sufficient money. These may reduce the number of test trials, the number of subjects, and the number of conditions tested.

3.4.2 The Checklist

One tool used frequently to evaluate design is the checklist — in written or mental form. The checklist is a series of statements describing the specific attributes that an equipment, procedure, or software program ought to have to be properly human engineered.

The design checklist is usually broken up for convenience into sections corresponding to major equipment subdivisions, such as scales, displays, etc.

Three aspects of checklist development are important to the measurement specialist, because there are relatively few standardized checklists:

(1) Determination of which human-machine interface (HMI) dimensions are important to operator performance. The interface has many dimensions, some which are important, others less so. The checklist developer must select those that are most important in terms of affecting operator performance.

(2) The dimensions selected must be expressed in such a way that the checklist user is quite definite about what they describe. The level of detail is critical.

(3) Checklist items may be expressed as positive statements requiring agreement or disagreement, or as questions, e.g., does the interface manifest the following characteristics?

The selection of equipment characteristics that might affect operator performance represents a judgment (deliberate or unconscious) of relative importance; this judgment may be in error. Checklist development, lacking empirical data on performance effects, is based largely on intuition.

The checklist represents a hierarchy of abstraction. For example, at the highest level is the attribute, e.g., operability. If one asks the question, “Is this equipment operable?”, it is impossible to make a meaningful evaluation, since operability is a function of a number of specific characteristics, each of which must be evaluated in its own right. One order of abstraction lower is the use of an evaluation criterion phrased in terms of a standard, such as Mil-Std-1472, e.g., are controls and displays that are functionally related to each other located in proximity to each other?

The evaluator uses a checklist item in two ways: (1) as a reminder that an evaluation of a particular characteristic is required (this is easy) and (2) as a means of making a comparison between the equipment and the attribute reflected in the checklist item (much less easy, as shown in Figure 2). If the checklist item is ambiguously written, the evaluator may have difficulty in constructing a mental picture of the characteristic. The greatest difficulty arises in judging whether the attribute being evaluated does or does not satisfy the standard implicit in the checklist item, because the item is usually phrased in dichotomous terms (yes or no), and the characteristic as one sees it is actually only one point on a continuum. A further deficiency is the lack of a method of deriving a quantitative figure of merit from the checklist. The checklist is an evaluative device but is used in system development mostly as a diagnostic tool to determine inadequacies that must be remedied.

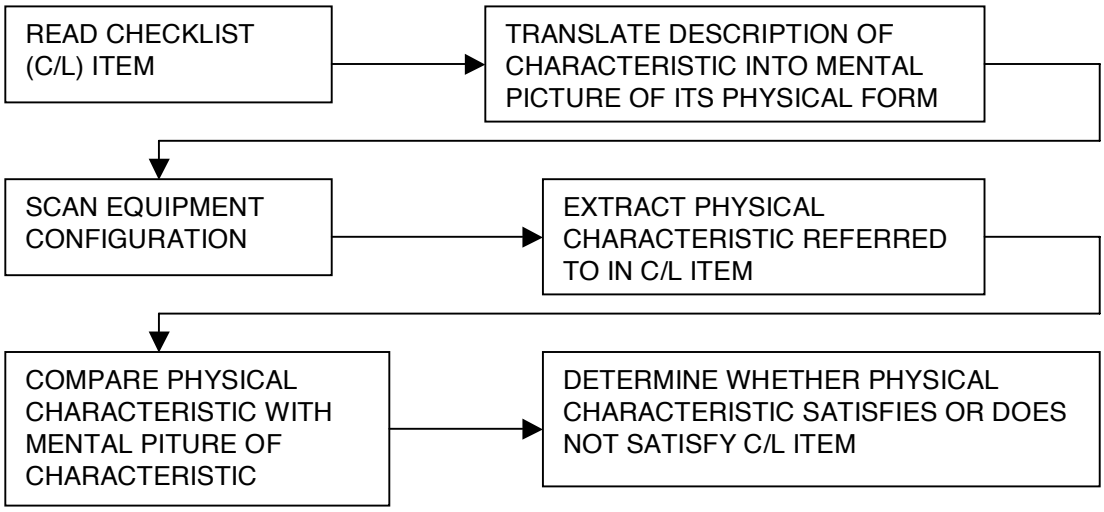


Figure 2 – The Checklist Comparison Process

A number of ergonomics checklists have been published for general use (e.g., Smith and Mosier (1986) and Williges and Williges (1984)).

3.4.3 Mockup Testing

Mockups are scaled representations of the physical characteristics of equipment and systems. The mockup in hardware development is essentially a tool to assist that development by enhancing conceptualization of the HMI and permitting limited equipment testing with subjects. Mockups may be two-dimensional, used for initial control panel layout studies; three-dimensional static; full-scale; and functional (i.e., interactive) full-scale. Our concern is with the functional full-scale mockup used for evaluation testing.

With the functional mockup, the evaluator is now able, for the first time in development, to evaluate the adequacy of the equipment's operating characteristics and procedures in relation to operator performance in a simulated operational situation.

If the functional mockup is sophisticated enough so that operational procedures can be performed in their entirety, the evaluator can approximate a true system test by determining whether the operator can perform the assigned functions, how long it takes to do them, the types of errors made and the error rate, and the problems the operator encounters. To the extent that the evaluator can vary equipment configurations, operating situations, or environments (e.g., terrain), it is now possible to apply experimental design procedures to the mockup test. Certain limitations of the mockup test will almost always be encountered, however.

Unless the mockup approximates a representation, it is a small stage on which to perform an evaluation. It is, however, tremendously important until development produces prototype hardware.

In hardware design, the walkthrough is an exercise in which an operating procedure is performed largely symbolically, using either a design drawing or a static mockup of equipment. The designer goes through each step of the procedure in sequence, pointing to the control that will be activated and the display that will be illuminated if the action were actually performed with equipment. The purpose of the walkthrough is to evaluate the adequacy, from a behavioral standpoint, of the procedure and the design it represents. No quantitative measures are taken, because both time and errors are irrelevancies (and, in any event, cannot be measured with such a procedure). There are only three questions to be considered in a walkthrough: (1) can the procedural steps be performed with the design as it is presently configured, and, (2) if not, what needs to be modified, and (3) what does the procedure demand of the operator?

The walkthrough process is essentially the same in software design, except that one is evaluating the adequacy of instructions to computers and, from a behavioral standpoint, what the instructions demand of the software user. The materials used may be flow charts of actions required by the software and the user, or a storyboard, where the user's actions and the resulting display on the CRT are considered.

The walkthrough is a conceptual evaluation and therefore, is inferior in evaluating performance-oriented measurement, such as rapid prototyping. Conceivably, it could have some value prior to rapid prototyping testing, but not afterward. On the other hand, if properly performed, rapid prototype testing is more expensive in money and time than the walkthrough, so some software developers may prefer the latter.

3.4.4 Rapid Prototyping

Rapid prototyping is the equivalent of mockup testing in software development.

Considerable mystique has grown up about this technique; it is considered one of the distinctive features of computer design. In essence, however, rapid prototyping is little different from the mockup testing of hardware design. Rapid prototyping consists of repeated testing early in design of new software, usually in the form of modules of the overall program. The testing is ideally performed with representative users,

sometimes with other software designers. Subjects are required to use the software to perform one or more tasks, and their performance of these tasks is observed and measured. Deficiencies in the new software are deduced from subject performance and then remedied; a new version of the software is developed and tested again by exposure to users. This procedure is followed until users are satisfied.

Every test of the software must involve a task to be performed. Presumably, this is a task that the software, when exercised by a user, has been designed to perform; that is, it should be an operational task. The assumption is that the software is inadequate if the software user in the test cannot perform the task, performs it with excessive difficulty, or requires an inordinate length of time to perform the test. The customer is always right; if the software does not permit the appropriately trained user to function effectively, it is unacceptable. This paradigm is essentially the same one that underlies the behavioral design/development of all machine devices.

Of course, this premise presumes that the operational task of the problem situation is known, and that the problem either replicates the operational task or is related to it. It would be strange and extremely inefficient if software were developed without a clear notion by the designer of what that software was supposed to accomplish. However, in very complex systems in which the software is designed in modules and then combined into an overall product, it is not uncommon for the ultimate operational tasks to be unclear in the designer's mind. That is why a function/task analysis phase is so important in software and hardware design.

If we focus on the task, we must also focus on the user. Problem situations should be attempted with potential users to determine whether the tasks to be performed are reasonable and do not make excessive demands on users. If not, they must be redesigned. Since the user's performance determines whether the software is adequate, the subject must be chosen with great care to be representative of the prospective user population.

If the test subject is unfamiliar with the software operation, he or she must be trained to at least minimal facility with the prototype software. The time it takes to sufficiently familiarize the subject in how to perform as a subject, and the difficulty encountered in learning, are additional criteria of software adequacy. If prolonged training is required, the software interface is probably too complex.

Several criteria of software adequacy exist. First and foremost is whether the subject can perform the problem tasks effectively. There may or may not be a maximum time constraint put on task performance, but, in any event, the designer should have some concept of what is a reasonable timeframe for task duration. Should even a few of the subjects take too long (even if they eventually perform the task correctly) the software is probably too complex.

The difficulties subjects experience, e.g., blind alleys, requests for aid from the designer, repeated resort to instruction manuals or on-line help provisions, also indicate that something is wrong with the program. All actions taken by subject users, and their verbal comments, should be recorded, because they may be indicative of problems the designer must overcome.

Most designers are familiar with the use of prototyping to determine the technical adequacy of the product they have designed. However, the methods used in rapid prototyping to determine behavioral adequacy are somewhat different. In rapid prototyping for behavior considerations, the user or someone truly representative of him or her must be a part of the prototype test.

To be truly effective, rapid prototyping must be conducted as all measurement should be conducted: formally, systematically, and efficiently. It is not acceptable to conduct rapid prototyping in an informal manner.

To be acceptable, rapid prototyping must include the development of test objectives; the specification of standards of satisfactory performance, and of criteria and measures; the selection of a sufficient number of representative users as subjects; a description of the training that subjects may require; and an

adequate data analysis following the test. These requirements are the hallmark of all behavioral measurement, and they should not be tampered with.

Harker (1988) presents general principles that apply to all prototype situations:

(1) *Adherence to design assumptions*. Prototypes of the human-computer interface (HCI) need not be based on the anticipated final form of the interface, but it is important that each prototype incorporate the same assumptions and constraints that will govern final design. The data collected from the prototype testing may (indeed, should) determine that final design; that is why the test is performed. Many situations exist where there is insufficient information to permit full simulation of the system; this constraint should be kept in mind when interpreting the test results.

(2) *Importance of "hands-on" use*. The prototype should be exercised in a realistic manner. Users acting as subjects should interact directly with the prototype, not merely passively inspecting the system or reacting to a demonstration of its characteristics by the designer. They should be tested for as long as necessary to allow them to become familiar and at ease with the system's operation. The amount of experience needed to produce meaningful data from a prototype test will vary with the individual system, but, if the system is large and complex and one is concerned with issues of organizational structure, any prototype evaluation may require days, weeks, or even months. Of course, one must distinguish between prototype tests conducted early in design and those conducted when the prototype is near completion; the earlier ones may be somewhat shorter than the later ones.

(3) *Sampling the user population*. Realism in the test requires recruitment of a representative sample of the projected user population. Harker points out that a major dimension affecting the determination of the user population is whether the system will be an "off-the-shelf" product at one end of the continuum, or tailored to a specific organization at the other end. Designers of off-the-shelf products face special difficulties, because their user population may be extremely diverse. In addition, they may be reluctant to take the prototype outside their facility for fear of risking disclosure of proprietary features. If the designer is developing a system for the company's own use or for a single organization, it should not be too difficult to recruit users from these organizations.

There is also the question of whether one wishes highly trained ("computer smart") users, which tend to make the prototype look good, because these users will experience fewer difficulties, or more naive users, whose difficulties will be useful in assessing training routines. This is a judgment call; the most desirable course of action is to have both in the subject sample and to do a differential analysis of the two types of users.

(4) *Ensuring the timeliness of prototyping*. To achieve maximum usefulness, the test must be conducted early enough in the design cycle so its data can be fed back in time to influence major decisions.

Particular attention should be paid to performance indices. These indices are: (1) performance on appropriate tests, (2) subject difficulties encountered, and (3) positive or negative attitudes expressed by subjects about the software stimuli. The specific data that must be gathered include: (1) time and errors in task performance; (2) percentages of problems solved or tasks successfully completed; (3) specification in the form of an interview or questionnaire of the difficulties encountered in using the software, including reference to the subject's knowledge of the program (mental model), which would indicate how well subjects had been trained to use the software; and (4) ratings or a checklist describing attitudes toward specific features of the software, such as feedback information provided in error messages. In some versions of the process, subjects might be asked to verbalize aloud what they are attempting to do at the time they do it.

Data analysis should be both quantitative and qualitative. This includes statistical comparison of subject performance with whatever the quantitative standard of acceptability is, analyses of the types of errors made, the time required to perform tasks, and content analysis of the interview/questionnaire and rating material in terms of design and training implications.

3.4.5 Verbal Protocol Methodology

This technique was developed by Bainbridge (1979). It requires subjects to verbalize what they are thinking as they perform tasks. Verbal protocol analysis has been used "to investigate cognitive processes involved in learning a programming language,...to examine interactions between different categories of behavior in mathematical problem solving,...and mental representations used by experts and novice mathematical problem solvers,...to explore differences between experts and novices solving political science problems,...to identify the learning procedures differentiating good from poor learners,...for eliciting knowledge from experts in order to construct 'expert systems',...examining sources of error,...[and] to identify the cognitive processes involved in carrying out a task" (Green, 1995, pp. 126-127).

Green (1995) has stated that there are eight distinct phases in using verbal protocol analysis: 1) task specification, 2) data collection, 3) data transcription, 4) exploration, 5) construction of a theoretical framework for the analysis of verbal data, 6) segmenting, 7) encoding, and 8) analysis. She also identified six classes of verbal reports: 1) concurrent (simultaneous reports), 2) retrospective, 3) individual reports, 4) group reports, 5) unaided verbalization, and 6) aided verbalization.

Although verbal protocol methodology is extremely intrusive, it has been used since no other technique can unambiguously identify information processing and decision making (Obata, et al., 1993). Robson and Crellin (1989) identified the separation of data and theory as one of the strengths of verbal protocol analysis. They point out, however, that the loss of nonverbal cues during the analysis can lead to misinterpretation. Brinkman (1993) has also raised methodological questions of reactivity ("the cognitive processes which normally proceed during task performance are changed by the mere requirement to verbalize them concurrently or retrospectively" (p. 1381) and invalidity.

Crutcher (1994) questions "whether verbal report data can be treated as objectively as other behavioral data" (p. 241). Payne (1994) states that "verbal protocol analysis can be exceedingly time-consuming" (p. 246). He suggests several analysis schemes, including frequency of occurrence of different types of reasoning and segmenting into unambiguous measures. Svenson (1983) recommended analyzing the verbal protocols by breaking them down into units of meaningful text and then analyzing the frequency of occurrence of each type of unit. Highhouse (1994) applied this analysis technique to evaluate an ambiguity model. Obata, et al. (1993) used a similar technique to evaluate three prototype In-Vehicle Navigation Systems.

Gilhooly and Green (1989) have developed a suite of computer programs, SPITBOL, to analyze protocols. Finally, Wilson (1994) stated "Verbal report methodologies cannot tap cognitive processes that never reach consciousness" (p. 251). The definitive text on protocol analysis is Ericsson and Simon (1993).

3.5 System Test and Evaluation

During the development of a system, testing is ongoing as detailed previously. Developmental testing and evaluation (DT&E) does not end with the hardware or software prototype, however. Operational test and evaluation (OT&E) starts at the concept exploration and definition phase, while developmental testing is paramount. As a program moves from the demonstration/validation phase to the engineering and manufacturing development phase, the DT&E tapers off and the OT&E becomes paramount. However, DT&E can continue throughout the life cycle of the program.

Prior to the development of a prototype, major modules of the system will have been tested individually, e.g., an aircraft's avionics, its navigational equipment, and its emergency backup equipment. Now, all these are assembled together, and the system is tested as a whole. The initial tests of the entire system are still considered developmental testing. For example, the Air Force Flight Test Center (AFFTC) conducts initial airworthiness, performance and flying qualities, avionics, subsystems, and human factors tests on all Air Force aircraft. This testing is considered DT&E. By its nature, human factors testing is very operationally oriented, with its concerns of operability and maintainability. The Air Force has directed

DT&E to become more operationally oriented, rather than only testing to specifications, and since initial OT&E is often concurrent to DT&E, the line separating them is often blurred.

Depending on the size of the system and the need for special test facilities, the OT&E conducted during the development phase can take place either in the engineering facility, at a special test range, or in the operational environment, e.g., an airport or the ocean.

Regardless of where the test is performed, one requirement is paramount: the system must be tested in a manner that reproduces the way it would ordinarily be employed operationally. If test conditions are artificial, the test will tell little or nothing about the system's eventual operational use. This means that, if the system has a specified mission, it must be exercised throughout that mission; or, if certain tasks are ordinarily performed with the system, those tasks must be performed as realistically as possible. This may create problems if the system is designed to be used in combat, but there are ways of approximating even combat.

It is also necessary that the personnel who will operate the system after it is accepted by the customer operate it during the OT&E. If operational personnel are not available for the test, then personnel essentially equivalent to them in training and skill level should be used as test subjects. It is not acceptable to use engineers who developed the system to act as test personnel during the OT&E, because their skill level is likely to be far greater than that of operational personnel. Under these circumstances, they would be less likely to make errors and more likely to work more efficiently than operational personnel. If one of the test purposes is to uncover weaknesses in the system resulting from behavioral inadequacies, use of engineering subjects would defeat that purpose. Even if operational personnel are used as test subjects, it is desirable that these contain a mix of high and low skill levels to approximate the range of capability found with operational personnel.

Should special training be required of test subjects to exercise the system effectively, that training must be provided and the subjects tested to ensure personnel adequacy.

Although the system may require a large test crew, it is not acceptable to have only one crew as test subjects because a single crew may, for unknown reasons, have certain peculiarities that render it atypical. A minimum of two crews is necessary, and more than two are desirable.

All operating and maintenance procedures should be completely formalized in writing because they are to be handed over to the eventual system users.

The test should be conducted in realistic operating conditions. For example, if the system must be operated in a sand storm or in turbulent water, it is necessary to simulate those conditions in the OT&E. All **routine** operating conditions must be replicated in the test. Emergency conditions (e.g., engine failure in an aircraft) may or may not be included as part of the OT&E, although it would clearly be desirable to do so.

The OT&E has a standard test cycle for all tests and should be adhered to rigorously. The test phases are as follows:

(1) *Write a test plan*. The plan is the formalization and end product of the planning process. Unless the test is planned, it will fail; unless the plan is written down, parts of it may be forgotten and may not be communicated to all who need to know about it. For further detail, see Lehman, *et al*. (1991).

The test plan objectives must be closely stated, measurable, and should evolve from users' critical issues as listed in the test and evaluation master plan (TEMP) or similar document. The plan should clearly state the evaluation criteria (i.e., pass/fail), the test methods, the instrumentation required, the data analysis plan, and the reporting plan.

(2) *Develop and try out data collection instruments.* There are practically no standardized (off-the-shelf) methods of collecting data unless one is using common timing and audio/video recording instruments. If observational and self-reporting techniques (e.g., observers, interviews, questionnaires, rating scales) are

used, they must be tailored to the special characteristics of the system (for guidelines, see Babbitt and Nystrom, 1989). In other words, if observation of personnel activities is to be used as a data collection technique, the things the observer should look for must be specified; if the observer is to be used to report errors, the nature of those errors must be indicated.

Although it is commonly thought that objective data collection methods and measures are superior to subjective ones, the evaluator will often find it necessary to use subjective ones in place of objective ones because of test constraints; in any event, even if there are objective instruments, the subjective ones often provide additional valuable information. For example, personnel who exercise the system should always be asked about their activities after the test is completed.

The use of subjective instruments makes a tryout of those instruments necessary, because non-standard instruments often contain flaws that are discovered only in use (e.g., interview questions whose meaning to the respondent is not quite clear). Such flaws must be discovered *prior* to the test. Even when only standard automatic instrumentation is employed, it is useful to have a preparatory run-through of the data collection process. For example, it may be found that a high ambient noise level requires adjusting voice recorders.

(3) *Try out total test procedure*. For all the above reasons, it is desirable to perform a complete test trial prior to commencing the actual test, so that any inadequacies can be discovered and remedied.

(4) *Run the test*.

(5) *Collect and analyze the data*. The operational test plan should contain a detailed method of data analysis, and the data should be analyzed according to that plan.

(6) *Write the test report*.

Since the operational test plan is the heart of the test, it requires some further description. The parts of the plan are:

(1) *Test purpose(s)*. The test will have at least one and perhaps more purposes, and it is necessary to describe these as clearly as possible. It is a tautology (and lacks measurement elegance) simply to state that the purpose of the test is to test the system. The main purpose of the test is to determine whether it is possible to perform such tasks in so many minutes or hours, with fewer than two errors per task, etc.

(2) *System description*. Specify exactly what will be tested. If the system is to be tested completely in all its operating modes, one need only refer to relevant operating manuals.

(3) *Experimental comparisons*. If the OT&E contains experimental comparisons such as night versus day operations, these comparisons and the reason for including them should be specified. The experimental comparisons may require a specific design (see 4.6), and, if it does, the impact of that design on test operations should be described.

(4) *Criteria and measures*. Measures are derived from criteria of what constitutes effective performance in the system being tested. There may be several criteria, and, hence, more than one measure may be applied. In any event, measures must be specified in detail so that appropriate data collection methods may be developed.

(5) *Data collection methods*. These are independent, in part, of the measures to be applied. The method to be used should be determined by its effectiveness in collecting the desired data and by the cost of using the method. The methods used should be described in detail so that all data collection personnel know exactly what they must do.

(6) *Subject characteristics*. The number and type of subject test personnel (those about whom behavioral data will be collected) should be specified.

(7) *Constraints*. The constraints under which behavioral data are collected may be so severe that they may reduce the data validity, their relevance, and the conclusions drawn from them. If this is the case,

the point must be made, so that management and the customer don't derive inappropriate and unwarranted conclusions.

(8) *Data analysis*. The analysis to be performed must be described in detail, because, if not, often much of the data collected cannot be properly analyzed (because statistical assumptions are inappropriate) or are irrelevant to the test purposes. This is particularly true if some sort of experimental design is to be applied to the test.

(9) *Test schedule*. If the test is to be run over long periods of time, or at irregular intervals, it must be determined if the schedule can be maintained, so that everyone knows the schedule. For further information, see Lehman, *et al*. (1991).

4 General Issues in Human Performance Measurement

Certain general issues underlie all HPMs. There may not be a great deal that one can do about these, but, to be effective in measurement and data interpretation, one must be aware of them.

4.1 Underlying Problems in Human Performance Measurement

Problems underlying all HPMs are:

(1) Lack of a general theory to guide performance measurement. If such a theoretical structure existed, it would relate behavioral processes within the individual to his or her performance of the task, and that task performance to total system performance. At present, only a few relationships have been discovered—for example, the inverse relationship between speed of task performance and accuracy, or performance quality.

(2) The inverse relationship between measurement control and operational realism. Control over the measurement situation is needed to manipulate experimental variables (see 4.6). One also wishes to maintain as much operational realism as possible, so that experimental conclusions can be generalized to the real world. In the real world, however, variables are not manipulated, and all variables are permitted to exercise an effect on task performance. The result is that there is constant tension between the need to control (with a consequent "clean" measurement situation) and the resultant artificiality of the experimental measurement setting. The laboratory, where the greatest amount of control can be exercised, often is accused of producing unrealistic results. Conversely, in most measurement cases, one wishes to determine the effect of certain variables on performance; this is difficult to do in a completely realistic situation.

(3) Behavior is multidimensional. Many factors influence human performance, some of great weight, others of little importance. Moreover, these factors probably change their importance over time and with different measurement contexts, and may have different weights in different individuals. This makes it necessary to perform more experiments with more variables, subjects, and test trials. The results of data measurement are also more obscure than would otherwise be the case.

(4) The relationship between objective and subjective data is unclear. Objective measures are based on observable behavior (i.e., moving a lever, speaking a word). Objective measures may be recorded by humans or machines; generally, machines are more accurate and reliable than humans for measuring simple, overt tasks. For complex tasks as well as phenomena that are not directly observable, however, objective measurements may not be possible or may be too costly to develop. Subjective data use human judgment to provide measures that may be less accurate and reliable than objective measures for simple tasks, but might offer more insight on human performance. Subjective data may be quantitative. For a complete review, see Hennessy (1990).

(5) Results may not be generalizable to the real world. For all the reasons addressed earlier, the meaning of data is often unclear. The overall meaning of experimental data is clear enough, because the meaning of the experimental variables gives meaning to the data. For example, if two training methods are experimentally compared and one is statistically significantly better than the other in terms of producing desired performance, the result can have occurred only because of the experimental condition

(assuming that the experiment was properly controlled). However, the generalizable meaning of the experimental data *in relation to the real world* may be unclear, because, in the latter, all factors (including the experimental ones) are operative; so, whether the factor producing the experimental result will now be effective operationally is undecided, unless we measure it again in the real world—which is not easy. On the other hand, measurement in a non-experimental situation does not permit one to examine the effects of individual variables.

(6) Measurement of cognitive tasks. The increase of computerization in task performance means that behavior is becoming more cognitive; personnel are acting as supervisors and managers of systems, rather than performing mechanical, physical operations that can be monitored easily. They respond only when the computerized system fails. Cognitive activity is inherently more difficult to measure than physical performance because it is within the individual. Cognitive activity cannot be observed directly; it requires analysis of the output consequences of the cognition and, even more important, some kind of self-reporting, which may be tainted by its subjectivity. This means that measurement of much behavior is becoming inherently more difficult than it was before.

In these cases, performance must be inferred from system inputs, outputs, and states, as well as from models of the operator's cognitive processes (Vreuls and Obermayer, 1985.)

(7) There is currently a lack of objective performance criteria for most tasks. This lack of criteria makes it difficult to (1) assess performance quality and sufficiency and (2) identify the type of performance measurement techniques required to operationally define significant differences in performance (Vreuls and Obermayer, 1985).

(8) It is extremely difficult to determine the contribution of each component being measured to the overall system performance (Meister, 1995).

4.2 Bridging the Gap Between Human and System Performance Measurement

4.2.1 Interactions of Human and System Performance and Their Exploration

The preceding subsection states that both human and system performance should be measured, because human performance always occurs in some sort of system context. If this is done, the question must be asked: how did human performance affect system performance and vice versa? It is assumed that, in a system directed by personnel, human performance will have some effect on system performance, but to what extent, and by what mechanisms is that effect produced?

These questions are not easy to answer, for several reasons: the relationship between human and system performance is rarely a linear one; a complex system has much inertia about its response to human actions; the redundancies and interdependencies found in complex systems dampen human performance effects; and extra-system factors may distort both system and human performance. For example, sonar efficiency (both human and system) in tracking submarines declines significantly if the submarine is traveling in a thermal current that sound waves cannot penetrate.

There are two ways of exploring the relationship between human and system outputs. One is quantitative; the other, qualitative. For some relationships neither may be conclusive and multiple tests may be required to establish the relationship.

The quantitative measures of human proficiency (e.g., speed of response, task completion time, task completion accuracy) can be correlated on successive trials with whatever the measure of system output is. To make the correlation work well, both the human and the system measures should vary; that is, they should not be binary, either a complete success or a complete failure.

The qualitative method of relating human to system performance involves conducting an audit of the effect of personnel failures, errors, delays, etc., on the individual system output. For example, if a pilot inadvertently switches his engine off, so that the plane accidentally loses altitude, a direct relationship between the human error and system performance can be deduced.

4.2.2 Developing Performance Criteria

Kliem (1982) developed a six-step process for developing performance criteria:

(1) perform a time study that identifies the input of, output of, and time to perform each task (i.e., a task analysis);

(2) identify the task objective and the best method to attain that objective;

(3) determine the best metric to quantify the output;

(4) record the standard practice;

(5) calculate the mean, median, and mode times to complete each task; and

(6) compute the criterion for each task.

For example,

(a) determine behaviors that will be observed,

(b) determine the number of people to be observed,

(c) develop a work sampling sheet,

(d) determine the time, frequency and number of observations, and

(e) calculate the criteria

$$\text{criterion} = \frac{\text{times in minutes}}{\text{number of work units}}$$

4.2.3 Using Expert Judgments of Task Performance

Task performance estimates based on expert judgments are often used to predict performance in systems that have not yet been built or in situations that have not yet been experienced. Knowles, et al. (1969) identified three problems associated with expert judgments: lack of validity, low reliability, and measurement error. They suggested a research approach to determining validity. To enhance reliability, they suggested a three-step approach: First, determine the dimensions that are used to make a judgment (e.g., task difficulty and severity of operational conditions). Second, identify the dimensions that are not being consistently judged. Third, remove systematic errors.

4.2.4 Modeling and Simulation

As computer aided design tools and digital models of humans and human performance become more prevalent, human factors engineers are applying these earlier and earlier in the design process. First, the physical interface is evaluated using electronic human models that predict the operator's ability to see and reach controls and displays. Next, cognitive models are used to help design the information interface and can predict performance times and accuracies. Finally, the manpower, personnel, and training requirements can be predicted using digital models originally developed for the United States Army and now extensively modified for civilian use. An excellent review of these issues is in Gawron, Dennison, and Biferno (1996).

4.3 Asking the Right Questions

Ultimately, the effectiveness of measurement depends on asking the right questions. The data gathering techniques used, the measures selected, and the method of analyzing the data are affected by the questions asked and, therefore, must match those questions.

The first question to be asked in developing a study is: does what I want to learn require experimental or non-experimental measurement? This depends on whether one wishes to determine the effect of

variables on performance. In general, for both basic and applied research, one does; in developmental and OST, again in general (but with occasional exceptions), one does not.

If the researcher decides to study variables, the next question is: which variables? The selection of variables can be made with the aid of a formal theory or by less formal deduction of the cause of a gap in available knowledge. Variables selected on the basis of a theory are more effective because the theory provides a framework for predicting which variables will be important in performance and in what way. It is often more efficient to test one specific variable at a time, rather than rely on a complex experimental design.

The questions must be detailed. It would be fruitless to research an area such as workload, merely as workload. What is it about workload that one wishes to know? This requires specifying variables. However, even when variables are not at issue, for example, in the OST, the specific aspects the investigator will measure may be unclear. The requirement for evaluation of a system is often phrased in terms of the need to verify or investigate abstract concepts, such as compatibility, operability, and intelligibility. These must be translated into specifics, because in the abstract they offer little practical guidance to the measurement specialist. For example, the general purpose of determining the operability of a sensor device requires that several more specific purposes be stated, e.g., to determine performance as a function of detection and classification range and target speed and size. These more specific purposes suggest certain measurement operations, such as the determination of detection accuracy, response time, and number of false alarms.

You also must specify what you intend to do with the results of the measurement. In the case of applied research, developmental testing, and OST testing, this is fairly clear cut, because the study is performed for a concrete purpose. This is not the case in basic research, where the results may or may not have useful applications. There is some controversy over whether basic research should have an application goal (in addition to the goal of furthering understanding). Basic research that cannot under any set of conceivable circumstances be used by someone other than the researcher is likely to be trivial and of little value. The effective researcher, therefore, will anticipate the results likely to be achieved and ask what the value of these results will be to someone other than the researcher or others in the same field. Such a consideration is likely to sharpen his or her measurement tools.

4.4 Selection Criteria for Human Performance Measures

Although the following criteria are important considerations in selecting a performance measure, no single measure or set of measures will have all these attributes. It is for the reader to judge which criteria are most important to his or her application. In addition, Kantowitz (1992) addressed the need for ecological validity (i.e., subject, variable, and setting representativeness) during system evaluation.

Baker and Salas (1992) provide six principles for selecting measures of team performance: 1) “For understanding teamwork, there is nothing more practical than a good theory;” 2) “What you see may not be what you get;” 3) “There is no escaping observation;” 4) “Applications, applications, applications;” 5) “Judges and measures must be reliable;” and 6) “Validation for practice and theory.”

Anastasi (1988) provided a suggested outline for test evaluation:

(A) General Information

- Title of test (including edition and forms if applicable).
- Author(s).
- Publisher, dates of publication, including dates of manuals, norms, and supplementary materials (especially important for tests whose content or norms may become outdated).
- Time to administer.

- Cost (booklets, answer sheets, other test materials, available scoring services).

(B) Brief Description of Purpose and Nature of Test

- General type of test (e.g., individual or group, performance, multiple aptitude battery, interest inventory).
- Population for which designed (age range, type of person).
- Nature of content (e.g., verbal, numerical, spatial, motor).
- Subtests and separate scores.
- Item types.

(C) Practical Evaluation

- Qualitative features of test materials (e.g., design of test booklet, editorial quality of content, ease of use, attractiveness, durability, appropriateness for test takers).
- Ease of administration, including facilities for computer administration.
- Clarity of directions.
- Scoring procedures, including computer-scoring services and available software.
- Examiner qualifications and training required.
- Face validity and test taker rapport.

(D) Technical Evaluation

(1) Norms

- Type (e.g., percentiles, standard scores).
- Standardization sample: nature, size, representativeness, procedures followed in obtaining sample, availability of subgroup norms (e.g., age, sex, education, occupation, region).

(2) Reliability

- Types and procedure (e.g., retest, parallel-form, split form, Kuder-Richardson or coefficient alpha), including size and nature of samples employed.
- Scorer reliability if applicable.
- Equivalence of forms.
- Long-term stability when available.

(3) Validity

- Appropriate types of validation procedures (e.g., content, criterion-related predictive or concurrent, construct).
- Specific procedures followed in assessing validity and results obtained.
- Size and nature of samples employed.

(E) Reviewer Comments

- From Mental Measurements Yearbooks and other sources.

(F) Summary Evaluation

- "Major strengths and weaknesses of the test, cutting across all parts of the outline." (pp. 676-677)

4.4.1 Appropriate Level of Detail

Measures should reflect the performance of interest with sufficient detail to permit a meaningful analysis. For example, if one is evaluating alternative control and display relationships, the performance of each step and control activation in a procedural sequence could be important in understanding the best configuration or potential for errors. On the other hand, collecting such detailed information might not be appropriate when comparing the effectiveness of two competing systems with dissimilar procedures; in this case, one should focus on measures of effectiveness (e.g., how well did the operator/maintainer and machine perform the intended purpose of the system?).

4.4.2 Reliability and Methods of Measurement

Reliability is used here to indicate the repeatability of a measure. If one measures the same behavior in exactly the same way under identical circumstances, it should result in the same value of the metric. In human performance measurement, however, individual differences among human operators, decision makers, and maintainers occur; even the same person may respond to successive trials differently because of learning or other effects. To adjust for this, the concept of reliability is extended from a value of a metric to a distribution of a metric; thus, if one obtains the same distribution with repeated measures, the metric is said to be reliable. Proposed metrics should be tested so that the degree of repeatability, often expressed as a correlation coefficient, is known. Lane (1986) stated that reliability is the most important criterion for selecting performance measures.

Murphy and Davidshofer (1991) identify four methods for measuring reliability. The first is the test-retest method. It involves: "(a) administering a test to a group of individuals; (b) re-administering that same test to the same group at some later time; and (c) correlating the first set of scores with the second." (pp. 78-79) This correlation is the estimate of the test reliability. The second method of estimating reliability is the alternate forms method. It involves: "(a) administering one form of the test (e.g., Form A) to a group of individuals; (b) at some later time, administering an alternate form of the test (e.g., Form B) to the same group of individuals; and (c) correlating scores on Form A with scores on Form B." (p. 80) This correlation is the test reliability. The third method, split-half, involves: "(a) administering a test to a group of individuals; (b) splitting the test in half; and (c) correlating scores on one-half of the test with scores on the other half. The correlation between these split halves is" (p. 81) the test reliability. The final method, internal consistency involves "(a) administering a test to a group of individuals; (b) computing the correlations among all items and computing the average of those correlations; and (c) using Formula [1] to estimate reliability." (p. 82)

$$r_{XX} = \frac{k(\overline{r_{ij}})}{1 + (k-1)\overline{r_{ij}}} \qquad [1]$$

where $\bar{r}_{ij}$ = average intercorrelation among test items

k = number of items in the test

Anastasi (1988) identified two additional methods for measuring reliability. The first is the Kuder-Richardson Reliability and Coefficient Alpha:

$$r = \left(\frac{n}{n-1}\right)\frac{SD^2 - \sum pq}{SD^2} \qquad \text{(p. 123)}$$

where r = reliability of the whole test

n = number of items in the test

SD = standard deviation of total scores on the test

p = proportion of persons who pass

q = proportion who do not pass each item

Anastasi's second method is scorer reliability. It "can be found by having a sample of test papers independently scored by two examiners. The two scores thus obtained by each test taker are then correlated in the usual way, and the resulting correlation coefficient is a measure of scorer reliability" (p. 125).

4.4.3 Validity

Does the measure mean what it is supposed to mean; is it appropriate to use for the intended purpose? There are at least six types of validity commonly distinguished: face, concurrent, content, construct, predictive, and criterion-related.

Face validity often involves the use of expert opinion to provide a starting point for developing human-system performance metrics. A subject matter expert usually confirms that the particular metric represents performance that is important for accomplishment of the task. Although face validity may not be sufficient to establish scientific proof that a measure is valid, it is important for user acceptance of results

Concurrent validity is the correlation of a measure with other measures. If two measures correlate highly with each other, they may be measuring the same thing. The higher the correlation, the greater the degree of similarity.

Content validity addresses comprehensiveness—proper sampling of the performance in a battery of test items and measures. Have you sampled all of the important areas of performance or knowledge? Do you have test items or measures that are unimportant, or perhaps irrelevant, to the task?

Construct validity is concerned with the correlation of a measure (or group of measures) with a construct, theory or model. One may hypothesize that responses (measures) to a written test battery (the measurement instrument) will be different for various professional groups, such as engineers, physicians, and artists. By offering the test battery to the various groups, one can classify the responses by group; if the responses are different, the validity of the construct would be demonstrated. Similarly, in the performance domain, one may hypothesize and validate the construct that expert operators perform in a different way than novices.

Predictive validity is perhaps the most important characteristic of behavioral measures, yet it often is the most neglected and difficult to obtain. Here, one would like to know that measures being taken in a laboratory, on a mockup, in a simulator, or during training are representative and predictive of the performance of the human being (and that system) in the real world—on the job.

Criterion-related validity is an estimate of "the effectiveness of a test in predicting an individual's performance in specified activities. For this purpose, performance on the test is checked against a criterion, that is, a direct and independent measure of that which the test is designed to predict" (Anastasi, 1988, p. 145).

4.4.4 Sensitivity

A measure must be sensitive to the behavior of interest. Response to pencil-and-paper test items is often substantially different in nature from the performance being studied, and must be shown to clearly tap the knowledge of interest. When measuring performance the proper measures must be selected. The measures must then be sampled often enough to capture the behavior of interest and be scaled for the desired range of behavior.

A common rule for continuous variables is to sample five (5) to seven (7) times more frequently than the highest expected frequency in that variable. Discrete variables, such as switch activations, must be sampled when they occur. The value that each measure can take should cover the full range that can be anticipated.

4.4.5 Diagnosticity

Diagnosticity can be thought of as the characteristic of a measure to provide information that will tend to isolate the cause of good or bad performance. Diagnostics may be used to improve system design or suggest ways to enhance human performance (or training). For example, the time required to finish a race may decide the winner (a "goodness" measure), but does not furnish information about why the winner won, or the losers lost; a measure of distance covered might provide diagnostic information, such as the winner covered less distance than the losers, and therefore traveled a more direct route. Measures that have diagnosticity add value to the measure set by providing information that might not be obtained in any other way. Lane (1986) lists three requirements for diagnosticity: 1) correct level of detail, 2) distinctiveness, and 3) easily mapped to components being evaluated.

4.4.6 Non-intrusiveness

A measure that attracts the attention of the subject during the process of data collection, may clearly affect the subject's task performance. If it does, the measure is intrusive. Note that it is not the measure that is intrusive, but rather the method of collecting the data.

Almost all measures are intrusive to a certain extent, but their contaminating effect on task performance will vary. Under certain, special circumstances, a data collection method will be completely non-intrusive, but these are rare. Examples are naturalistic observation when the observer is invisible to the subjects, as with a one-way vision screen; or data are collected automatically, so the subject is not aware of data collection, as in the case of computerized recording of key strokes.

Less obtrusive methods of data collection are preferred over the more intrusive ones. An example of a highly intrusive method is measurement of eye fixations, because this requires a camera attached to the eye.

In most test situations, the subject is aware that his or her performance is being measured. What the measurement specialist wants to know is whether this awareness significantly affects what the subject does. This is, of course, a "judgment call."

To the extent that the data collection method does not significantly alter the context of the task being performed, the method will probably not unduly affect the subject. The worst case, of course, is one in which the subject is forced to alter task performance to provide data, e.g., switching task performance to filling out a data form; this is insupportable from a measurement standpoint.

Self-reporting measurement techniques, such as interviews and questionnaires that can be used *after* task performance, are not intrusive. Automatic data recording, such as audio recording and timing, also are not intrusive. Video camera recording, if performed quietly, is only slightly intrusive.

When in doubt about task performance intrusion, the investigator should consider running a preliminary study using the data collection method. This might involve collecting a data sample and asking subjects to rate intrusion, or comparing monitored performance against unmonitored performance using an independent measure.

The selection of a data collection method is, after all, a tradeoff among competing requirements and constraints, and one may be forced against one's wishes to use a more intrusive methodology than is desirable. In making such tradeoffs, the investigator should ask the following questions:

(1) Will the subject be aware of the measuring device and the fact that performance is being measured?

(2) If so, to what extent will the subject be aware of this?

(3) Is it reasonable to assume that performance will be significantly affected as a consequence of the data collection method?

4.4.7 Implementation Requirements

When designing measures or selecting methods of measuring, one must consider the implementation requirements of the measuring set, and design the measurement system (be it manual or automatic), and determine what it will take in time, budget, personnel (training and operation), supplies, equipment, modifications to existing equipment, and logistics, to implement and maintain the desired measures. These considerations must include data quality control, storage, reduction, and analysis. Further discussions of these issues can be found in Swink, et al. (1978); Obermayer, et. al (1974); and Eggemeier, et al. (1989).

The operational practicality of the entire measurement system must be ensured when designing measures and measurement instruments. Issues to be considered include ease of data collection, robustness of the measurement instruments, and overall data quality control. These issues apply to the design and maintenance of laboratory measurement systems and instruments as well as to simulator studies and field exercises, but they are particularly important in field studies.

The ease of data collection includes devising practical ways to observe the performance of interest; the design of data collection instruments so they are self-evident and easy to use; and, if instrumentation is required, the development of sensors and instrumentation that are easy to install, monitor, and disassemble.

The robustness of measurement instruments, sensors, and instrumentation requires special attention if the data collection is to be in a field (or any non-laboratory) environment. Conditions in the field and even in some simulators are less than ideal for measurement; weather, motion, vibration, and handling effects must be considered; and measurement instruments, sensors and instrumentation should be designed to withstand these factors. Also, power, heat, and shelter may not be provided by the operational command; portable facilities may need to be part of the measurement plan.

Finally, the issue of overall data collection quality control must be addressed, particularly for field studies. Investigators must develop practical ways to oversee the quality of data collection—to ensure that the proper data are collected at the proper time, and that compromised data do not enter the study analysis database.

4.4.8 Operator Acceptance

Operators, training personnel, and subject matter experts must accept the measures that are taken (as discussed under the topic Face Validity), or the analyst will face an uphill battle gaining acceptance of the study's results or training by end users. Operators and end users must believe that the measures cover the important aspects of the task(s) at hand or the requirements. Frequently, one must be able to trace each measure to a performance or training requirement to satisfy such needs.

4.4.9 Fairness

The data collected must describe a fair sampling or representation of performance or opinion. According to Edwards and Verdini (1986), fairness means equity. The measures must be capable of defining "good" as well as "bad" performance, and must explore all relevant features of competing systems. Investigators must guard against inadvertently biasing the results of a study through the selection or design of measures.

4.4.10 Accuracy

There is more to accuracy than simple precision of measurement. "Accuracy means precision, reliability and minimization of measurement error" (Edwards and Verdini, 1986). Measurement instruments must

be scaled and tested to ensure sufficient accuracy and inter-rater reliability. Electronic instrumentation also must be scaled, calibrated, and checked periodically to ensure that accuracy is maintained; typically, calibration test signals must be produced to ensure system integrity. At the same time, the measures need be no more precise than the phenomena being measured (for example, measuring human reaction time in nanoseconds would be meaningless).

4.4.11 Simplicity

Using simple measures is desirable, but doesn't mean that measuring complex tasks or operations must be simple when the task is complex; it does mean that one should strive to use the simple rather than the complex, whenever possible. For example, will average error suffice in place of root-mean-square (rms) error? Will quartile ranges suffice for variability? Will a discrete value suffice for a continuous variable? Edwards and Verdini (1986) offer guidance that a simple Performance Measurement System (PMS) minimizes length and maximizes understandability.

4.4.12 Timeliness

It is important that performance data be provided at the proper time. For training, the proper time is when the performance feedback will provide the most training value—usually, this is during or soon after the completion of a training task or exercise. In laboratory and field studies, it is common practice to collect data, then reduce the data when there is time. This can be a problematical practice, because it denies early detection of experimental problems or trends; moreover, if one uses modern screening or fractional factorial types of experimental designs, performance data are needed at the end of each task to decide where (in performance space) to collect the next data points. The value of performance data often decays with passing time.

4.4.13 Objectivity

Objective measures are impersonal; they exist external to the observer and give the facts as they are, without bias. This is opposed to subjective measures, which are inward and usually concern the thoughts, feelings, or opinions of the individual who is reporting the measures. Objective measures of human-system performance usually represent observable behavior. It should be the goal of all HPMs to record objective measures. However, subjective measures are often useful in interpreting the objective data.

4.4.14 Quantitativeness/Qualitativeness

Quantitativeness is concerned with quantity, or amount—how much there is of a substance, action, or whatever is being measured. Qualitativeness would be the kind of thing it is—or whatever is special about an object that makes it what it is. It should be the goal of all HPMs to record quantitative measures. Note that quantitative does not preclude subjective measures.

4.4.15 Monetary Cost

There are two types of cost: (1) recurring and (2) non-recurring. Recurring costs are those that continue to be incurred during the life cycle of the product. Recurring costs of HPM include items such as data collection, materials, and computer use. Nonrecurring costs are those that may occur only once in the development cycle of a HPM. The nonrecurring costs of manufacturing include things such as initial engineering design and validation.

4.4.16 Flexibility

Measurement instruments and automated performance measurement systems should be designed in a manner that will enhance the ability to make changes in measures as situations demand. For automated, computer-driven performance measurement systems, this means placing measurement specifications in tables (or other such mechanisms) so that changes can be made without having to recode or recompile the computer program.

4.4.17 Utility

One should assess the value of acquiring measures versus the cost of doing so, as discussed in Finley, et al. (1975).

4.5 Measurement Uncertainty

Analyzing HPM results requires knowledge of combining errors, outlying data, and the speed-accuracy tradeoff.

4.5.1 Combining Errors

Errors in sampling, test-item construction, and measurement may coalesce in unwanted ways. Errors can be additive, making performance appear to be worse than it really is, or they can cancel each other, making performance appear to be good when it is bad. Also, errors may be made when combining measures into a more molar score—unlike measures may be grouped, correlated measures may be added together, or an unvalidated weighting scheme may be used.

The best protection is to measure at the most microscopic level possible when constructing the measurement system, then pretest the measures. During the pretest, examine the characteristics, distributions, and correlations of all measures to be sure of what you have before combining measures or measuring at a more molar level of performance. In addition, one should devise methods for independent assessments or quality checks on performance; exercises required to demonstrate measurement validity will go a long way toward reducing error. See also 4.6. Williges, et al. (1992) have developed guidelines for combining data from sequential experiments. Their guidelines were developed to minimize combining errors.

4.5.2 Outliers

Aberrant performance may be found in many data collection situations. Such performance may be well outside the expected range or the range of most existing samples of data. Usually, there are reasons for such performance. The experimental design, controls, sampling equipment, scenario, subject motivation or instructions might be faulty; a subject may be temporarily ill or emotionally upset. If a reason can be found that indicates that the data are not representative of the rest of the data set, the data should be removed. Usually, all data for that subject are then removed, depending on the experimental design. If the data are retained, nonparametric statistical analyses should be used to avoid the effect of outliers on the parametric data distribution.

This issue is controversial, however, removal of outliers depends on the data collection situation and use of the data in question. A sample that is three or four standard deviations above the mean may or may not bias the analysis of single measures significantly but can wreak havoc on multivariate procedures, such as regression or multiple-regression analyses. The investigator should examine the data with and without outliers and make a professional judgment. If outliers are removed, the same rules must be used to remove all outliers, and the issue and procedures should be documented. If possible, the results should be validated with a new sample. See 4.6.

4.5.3 Speed/Accuracy Tradeoff

Many human performance tests include accuracy and speed as separate measures, but investigators should bear in mind that humans may trade speed for accuracy, or the converse, and that the strategies for doing so may shift during the measurement interval. Speed and accuracy tend to be inversely related, but this is not always the case, and the relationship may not be linear.

Investigators must remember this aspect of human behavior. Where this might be a problem, explicit instructions (maximize speed, make as few errors as possible, etc.) are required. See 4.6.

4.5.4 Biases in Usability Measurement

There are fundamental differences between usability studies and experiments. The purpose of a usability study is to measure a construct; an experiment tests a hypothesis. It is the responsibility of those proposing measurement tools to show *construct* validity (demonstrate that usability measurements do indeed evaluate "usability"), *content* validity (show that each metric relates to usability), and *predictive or concurrent* validity (demonstrate how well usability measurements actually relate to customers' assessments of product usability) Cronbach (1970).

In experiments all methodology, materials, and procedures can be precisely defined, allowing others to replicate the experiment. In usability studies, it is just not possible to define all of the variables that can and do affect their use as a measurement tool of product usability. These variables or biases, such as the selection of tasks, subjects, experimenter, instructions, procedures, and reporting, have an undesirable effect upon measurements. These biases bring into question whether different experimenters using the same measures will arrive at the same numbers when evaluating the same application.

4.5.4.1 Specific Biases

Presented below are brief discussions of at least some of these biases.

Selection of Task Bias: Human factors practitioners are biased in choosing which product tasks to evaluate. We select tasks to evaluate what we believe to be representative of what our users will do with the product and, more revealing, tasks that are manageable and suitable for a lab evaluation. Tasks we typically target for evaluation:

(1) *Last 4 hours or less.* Long tasks reduce the number of subjects that we can run. Plus, there are problems with subject availability and attrition.

(2) *Tasks we know how to do.* We would never want to be in the position of not knowing how to perform a task we are asking our subjects to do.

(3) *Tasks that can be set up in our labs.* We are generally reluctant to conduct evaluations outside of our controlled environments.

(4) *Tasks that we find interesting.* Human factors people are biased against testing some user interface areas, such as documentation and error messages in favor of the more glamorous software menu or window interface.

(5) *Tasks for which we can find subjects.* We like to be able to run representative or actual users on our products; their availability affects the makeup of the tasks we choose to evaluate.

Selection of Subject Bias: Again, we like to run representative or actual users on our products. However, many studies are done with company volunteers or paid temporaries from an employment agency. Some biases with this subject pool are:

(1) *Volunteer bias.* Do you ever ask yourself why internal volunteers seem to have the free time to be able to participate in your study? Most likely (but not always) they are not the key people in their areas, and there may be good reason for that.

(2) *Company volunteers are not typical.* The people in your company are probably intelligent, inquisitive, and like solving problems. This may not be an accurate description of some of your customers, who may not like the challenge of learning an application.

(3) *I want a job bias.* Often, temporary people from employment agencies are used as subjects. Many of these people feel that if they do well to please you they might have a chance of being hired. Accordingly, they may mask negative feelings they may have about your product so as not to offend.

(4) *Age, sex, and seasonal bias.* Most temporaries are young females. In the summer, they are mostly college students; in the fall, mostly folks with kids in school; in winter/spring, you find that more chronically under-employed are available.

Experimenter Bias: It has been shown that the gender of the experimenters can affect the results of experiments, Rosenthal (1967), as well as their expectations, Rosenthal (1966). It is striking to watch how the individual style or personality of experimenters can affect subject performance and attitude. An experimenter's empathy, detachment, and tolerance are hard to quantify, but they can and do affect the outcome of usability studies.

Instructional Bias: How we instruct the subjects in a usability evaluation can have a profound effect on the results. In a study evaluating subject attitude bias, Cordes (1989a) found that participants in a usability study gave up over six times more often when a single sentence was added to the instruction that questioned their belief that all the tasks could be performed. Hutson (1989) has shown that instructions leading a person to believe they are responsible for their performance can have a dramatic effect on whether they report a task as being difficult.

Procedural Bias: So much of what makes up the procedural rules of usability evaluations is arbitrary. For example, how long do you let subjects "sweat" until you count it as a failure? If they fail, do you give subjects the answer and then move on? How much non-task related assistance do you provide (e.g., instructional clarification and keyboard assistance)? When is a mistake considered a real error worth recording? If they get the right answer, but don't realize it, does it count? Does the clock start running when they begin using the product or when they spend a half hour reading the user's guide? Do you count the time subjects may spend explaining problems or asking questions about your product? If a subject gets too frustrated and wants to quit, do you count that and all subsequent tasks as a failure? At what point do you think a frustrated subject has had enough?

The answer to many of these questions is that it is a judgment call of the experimenter. Since different people will make these judgments differently, variability among experimenters is inevitable. For instance, Hutson (1989) has shown that such a simple variable as the timing of when you collect subjective measures can have an impact on subjective ratings.

Reporting Bias: In any usability evaluation, there are typically hundreds of results and numbers that can be reported. We selectively go through this wealth of information and report only that which we think is important. What we think is important is affected or clouded by that which we understand (who wants to report an enigma?), that which we did correctly (no one likes to highlight their dirty laundry), that which we think would be important to our customers, and that which underscores a position we hold dearly. Indeed, it is a truism that, whenever we go through the trouble of conducting a usability evaluation, we will always find problems with a product; since we discovered them, these problems will be of utmost importance to us to have fixed. The fact that most practitioners will always find problems that they feel are important reflects a possible lack of objectivity and perspective.

4.5.4.2 Effects of Biases on Usability Measurement

Can't Rely on Metrics—Not Repeatable: Clearly, if the results of an evaluation are dependent upon the decisions, biases, and ability of the experimenter, then the numbers produced may be of little use. Because of the biases previously listed, there is a low probability that the measurements resulting from one evaluation can be independently verified by another investigator. Too many of the measurements obtained are the result of the tasks chosen to evaluate, the subjects selected to participate, the personality of the experimenter, etc. For example, how likely would it be that two experimenters independently, without collaboration, would choose the same tasks to evaluate, the same or similar subjects, or even the same rules for conducting their evaluations? Even though they may have the same measurements to collect or even the same rules, all of the other myriad of unspecifiable decisions that make up an evaluation will have a considerable and indeterminable influence on those measures.

Can't Generalize Results to Customers: A laboratory is a very artificial environment in which to attempt to evaluate product usability. For example, people asked to perform product tasks in a laboratory typically assume that each task you ask them to do can be done with your product. It is simply part of the implicit demand characteristics (Orne, 1962) of being in a study. Part of using a product is learning what it can or cannot do. In usability evaluations, our subjects do not have to make this determination. This

makes the laboratory situation far removed from what our customers experience. Indeed, it was shown, in a double-blind usability study (Cordes, 1989b), that participants receiving normal instructions were only willing to give up on tasks three times. However, in another group in which they were not 100% sure that the tasks could be done, the participants made a total of 19 "I give up" calls. The latter group more truly represents a customer's situation. Most usability evaluations succumb to this bias and, therefore, under-represent the severity of the usability problems encountered.

Another difference between a laboratory situation and that of the users is that subjects in the lab are typically alone, with the product documentation conveniently within reach. In the "real world," this is rarely the case. Users of products seek out other, hopefully more experienced, users whenever they need help. They live and work in a community, not in isolation. Also, they rarely have the documentation available to them at their workstation or office. Typically, they either have to borrow it from someone else or "wing it" with help from a few friends. Having the "latest release" of books is something of a luxury, and trying to match software releases with documentation can be a nightmare.

4.5.4.3 Test—But Be Aware of Biases

The fact that usability testing is biased in nature doesn't mean that testing shouldn't be performed. Performing frequent usability testing will probably improve the usability of products—because of the problems found and corrected, not because of the numbers that can be produced. Usability testing results in people uncovering usability problems and is a vital requirement for improving products. The key is to be aware of these biases and to be wary of the measurements that can result from these studies. This is especially true if they are treated as absolute indicators of product usability. On the other hand, relative measures comparing a product with its previous release or with that of a competitor may have some meaning (mainly direction)—if the same investigators use the same tasks with similar subjects. The measures obtained in iterative testing methodologies (test, make changes, and test again) can be quite useful in steering the development of a product into a more usable direction.

Still, you must make sure that the measures taken do indeed reflect those aspects of a product that the users will notice and find important. For example, it does little good improving on a measure twofold if nobody notices. Indeed, it may be best to focus your measurements and improvements in those areas where users appear to have the most difficulty. Improving other, less problematic, areas may not only be ineffective, but also may actually be detrimental to overall perceptions of usability (Cordes, 1993).

For now, it is best to admit to the subjective nature of usability measurement and do more work closing the loop between customer-perceived usability and lab studies. With this information, we may be able to determine what actually does allow us to improve and predict a product's usability. Without this validation, we will simply never know the answer. It is like throwing darts at a board—blindfolded.

4.6 Experimental Design

No measurement, basic or applied, should be performed without knowledge of basic experimental design principles. The reader is referred to Stanley and Campbell (1966) and Van Cott and Chapanis (1972).

For guidance in determining the appropriate number of subjects, see Virzi (1992).

4.7 Instrumentation

The paragraphs below provide guidance in the selection and application of tools associated with HPM. Additional guidance is provided in 2.8. The format established by the American Psychological Association (APA, 1983) requires that the instruments used in an evaluation be reported with the results and conclusions. The level of detail should be sufficient to enable the reader to obtain comparable equipment. Commercially marketed instruments should be accompanied by the manufacturer's name and model number. In the case of methodological instruments, such as a workload rating scale, the reader should be given a publication reference for further information. If mockups or other custom-made

equipment are used, the material, dimensions, and cross-reference to blueprints, drawings, or sketches should be included. The types of HPM instrumentation are summarized in Figure 3.

HP Test & Evaluation Method	SELECTION EVALUATION CHARACTERISTIC															
	Most applicable program phase				Relative complexity			Relative time to perform			Relative cost			Relative cost effectiveness		
	Concept	Program definition	Engineering and manufacturing development	Prod/Dev/Ops	Simple	Average	Complex	Short	Medium	Long	Low	Medium	High	Low	Medium	High
1 Continuous direct observation		X	X		X				X			X			X	
2 Sampled direct observation		X	X		X			X			X				X	
3 Specification compliance summary sheet		X	X		X				X			X			X	
4 Technical manual function evaluation		X	X	X	X				X			X			X	
5 Human Factors Engineering Data Guide for Evaluation		X	X	X	X				X	X		X			X	X
6 Environment and engineering measurement equipment		X	X	X		X		X	X			X			X	
7 Systems records review		X	X	X	X				X	X		X			X	
8 Test participant history record		X	X	X	X			X			X				X	
9 Interview	X	X	X	X		X			X			X				X
10 Questionnaire	X	X	X	X	X				X			X				X
11 Motion pictures		X	X			X			X				X	X	X	
12 Sound recording		X	X		X			X			X			X		
13 Video recording		X	X			X			X				X		X	
14 Still photography		X	X	X	X			X			X				X	
15 Event recording		X	X			X		X	X		X	X			X	
16 Secondary task monitoring			X				X		X	X			X	X		
17 Physiological measurement			X				X		X	X			X	X		
18 Physical measurement		X	X			X			X			X			X	
19 Online interactive simulation		X	X				X			X			X	X	X	
20 Statistical analysis		X	X			X	X		X	X		X			X	
Note: Information for the above items was collected in 1986 and may be dated or not applicable if automated techniques are used.																

Figure 3 – HPM Method selection (Mil-Hdbk-46855A)

4.7.1 Accelerometer

An accelerometer is a device, such as a strain gauge or piezoelectric force transducer, that measures acceleration along one or more axes. Accelerometers are used to measure the forces imposed upon the subject during the performance measurement period; they are particularly important in aeronautical applications.

4.7.2 Anemometer

An anemometer measures local air flow in the range of 0 to approximately 350 meters per minute. It is useful for determining whether sufficient air flow is provided for sustaining adequate human performance (Mil-Hdbk-46855A). The anemometer is used to determine the adjustments to a sound pressure level (SPL) when making outside sound measurements.

4.7.3 Anthropometry Instrument Kit

This kit allows measurement of significant body dimensions using an anthropometer, spreading calipers, sliding caliper, goniometer, and tape measure. These instruments are used to measure subject size and motion, particularly when evaluating the relationship between human performance and the physical layout of workstations. Care should be taken to ensure that proper measurement procedures are adhered to while obtaining participant anthropometric data (Mil-Hdbk-46855A). Proper training is required.

4.7.4 Electrogoniometer

An electrogoniometer is an electric goniometer—that is, a device (usually a potentiometer) that measures a change in joint angle as a voltage difference. Like the goniometer, the electrogoniometer is used to measure subject motion, typically in evaluating the impact of workstation design on operator performance.

4.7.5 Force, Torque, and Dimension Kit

This kit includes various instruments for measurement of a wide variety of operator or equipment forces, torques, and distances. Force measurement limits should be from approximately 7 grams to 120 kilograms. Torque measurement should range from approximately 0.05 to 250 newton-meters. The tape measure should be capable of measuring distances up to 15 meters. Scales should also be provided for measuring centimeters, millimeters, inches, and fractions of inches. A protractor is useful for angular measurement (Mil-Hdbk-46855A). These instruments are used to quantitatively describe the workstation and equipment, and to measure the forces and torques provided by the operator.

4.7.6 Gas Tester

A gas tester permits convenient short-term sampling and evaluation of many toxic gases, vapors, and fumes (Mil-Hdbk-46855A). It is used to determine the presence and concentration of toxic airborne elements that may affect human performance. There are, however, other ways of sampling for toxic gases. Consult an industrial hygienist for the best method for a given gas.

4.7.7 Hygrometer or Psychrometer

This device measures relative humidity by the wet and dry bulb thermometer method (Mil-Hdbk-46855A). It is used to measure the amount of moisture in the air, a factor that may affect human performance over extended periods of time. The hygrometer or psychrometer is used to adjust SPL measurements based on the water content contribution to air density.

4.7.8 Motion Measurement Systems

Current motion measurement systems digitize video data for subsequent analyses. These analyses include calculating accelerations, angles, angular velocities, displacements, total body centers of mass, and velocities. Some systems, such as the Peak 2-D video/computer motion measurement system, even have the capability of building spatial models from the digitized data. Typical systems provide 0.5-percent

resolution, color graphics, and computation of 3-D coordinates. Such systems are useful in gathering and analyzing data from time-and-motion studies of human performance.

4.7.9 Photometer

A photometer measures ambient illumination over a range of levels from approximately 0.05 to 270,000 lux. It is particularly valuable for verifying specification compliance with light-level requirements and in determining the impact of lighting conditions on operator performance (Mil-Hdbk-46855A).

4.7.10 Sound Level Meter and Analyzer

This device measures steady-state sound in the approximate range from 10 to 150 dB for standard weighted noise curves. The analyzer provides octave-band analysis for the more critical speech-range center frequencies. Specification compliance in terms of noise curves and speech interference levels may be verified with this equipment. Hazards to test personnel may be checked prior to overexposure conditions (Mil-Hdbk-46855A).

4.7.11 Spot Brightness Meter

A spot brightness meter measures luminance in candela per square meters for small areas. It is most useful for measuring prototype hardware display brightness, such as from LEDs or CRTs, to determine whether brightness enables adequate human performance (Mil-Hdbk-46855A).

4.7.12 Thermometer

A thermometer measures air, surface, or liquid temperatures. It may provide a digital readout in either Celsius (centigrade), Fahrenheit, or Wet-Bulb Globe Temperature (WBGT). A thermometer should have the capability for attachment to several temperature-sensor probes (Mil-Hdbk-46855A). The temperature of the workspace can significantly impact human performance and is measured to confirm compliance with specification, or to assess its impact on measured performance.

4.7.13 Vibration Meter and Analyzer

This device measures amplitude and frequency components of complex vibrations. The analyzer may be used to determine amplitudes at selectable frequency bands in a range from 2.5 Hz to 25 kHz. Potential vibration hazards to test participants may be checked before actual test exposure, and the effects of vibration on human performance may be quantitatively determined (Mil-Hdbk-46855A).

4.7.14 Video Tape Instrumentation

Video tapes are becoming an increasingly popular instrument for the gathering of human performance data. They provide the data for digitized motion measurement systems (see 4.7.8) and provide excellent records of visual and audio operator performance for post-trial analysis. A complete system comprises a television camera, a video tape recorder (VTR), and a television monitor. A number of guidelines have been prepared to promote the effective use of this technique (Crites, 1980):

(1) Complex tasks should be recorded simultaneously from different aspect angles.

(2) Zoom lenses with focal-length ratios of at least 8:1 should be used. This allows the framing of the subject without having to move the camera. Remote-controlled focus and zoom are recommended. A wide-angle lens of 5 mm or higher is recommended for work inside compartments.

(3) Small hand-held cameras are necessary for use in cockpits and compartments.

(4) A video/audio junction box facilitates the selection of the audio and video inputs for recording.

(5) The VTR should feature a conventional counter to locate scenes on the tape. A time display should also be included to allow accurate timing of performance.

(6) Ample lighting is required for good detail, depth of focus, form, and shape. Even common floodlights and spotlights may not offer the proper characteristics.

(7) Cameras should have good low-light-level imaging characteristics.

If the VTR does not have a time display or a more accurate time display is needed, a time code generator can add a time code on the video tape with a resolution of 1/1000 of a second.

4.7.15 Spectroradiometer

This device measures the radiant energy in the ultraviolet, visible, and/or infrared spectrum. Ranges are typically 200 to 780, 380 to 780, or 380 to 1070 nanometers. The spectroradiometer provides color coordinates, transmissivity, and spectral range information. Specification compliance to color standards, Fed-Std-595, may be verified with this equipment. It can check whether lighting is compatible with night vision goggles. It can also be used to identify the extent to which colored indicators can be seen through colored filters or in colored ambient lighting.

4.7.16 Video Digitizer

Video digitizers are typically add-in boards for computers that digitize a frame of composite or SVHS video for processing in the computer. Resolutions range from 250 horizontal pixels by 250 vertical pixels to 1024 horizontal pixels by 1024 vertical pixels. There are many uses for video digitizers, including motion studies, anthropometric joint angle measurement, smoke elimination rate measurement, engine exhaust visibility measurement, remote distance measurement, remote angle measurement, image enhancement, and image manipulation. Video digitizers allow computer processing of video tape data.

4.7.17 Digital Audio Tape (DAT) Recorder

DAT recorders can be used to record sound measurements up to 110 dB for later analysis. Signal-to-noise ratios for DAT recorders are in excess of 70 dB, which makes them ideally suited for high quality sound measurements such as recording ambient noise levels, tonal annunciators, verbal annunciators, communication line losses, and radio transmission losses. A DAT recorder may also be used as the “speaker” for consistent pronunciation reproduction of word sets for articulation index studies.

4.8 Data Collection and Analysis

4.8.1 General Considerations

Data collection and analysis techniques play a crucial role in the evaluation of human performance. If the human factors engineer has questions concerning data collection, analysis, or interpretation, a statistical specialist should be consulted. This consultation should occur during the early planning stages, as errors in sample selection or data collection procedures can seldom be corrected during data analysis.

The data collection and reduction processes should be pretested before beginning the formal testing process, particularly when test time is at a premium, to ensure that the equipment and procedures work as expected. Sufficient lead time should exist to fix non-catastrophic problems or to correct procedures.

4.8.2 Data Collection

Table 1 offers guidelines for the standardization of data collection media and media formats. Compliance with common industry standards permits the use of a variety of data storage and analysis tools. The most important issue is not with data collection but with archiving the collected data for record and future use, such as re-analysis or baselines for system comparisons. The high storage capacity of computer media permits archiving the data in alternative formats to maximize its future use. Additionally, picture files may be preserved in both compressed and uncompressed formats. Standardization enables the development and maintenance of databases of human performance data having the broadest range of potential users and potential archival longevity of hundreds of years.

Recognizing the realities of working in the field, the closer one comes to a laboratory setting, the more sophisticated the data collection can be. Whenever possible, collect quantitative data in the field in computer-compatible format. All collected data should be converted to standard computer file formats. The CD-ROM is an excellent medium for data storage, replication, and use by almost every desktop computer system. The CD-ROM has the virtue of having excellent archival quality (100 to 200 years) as well as being capable of being read on almost all computer systems. The best quality CD-ROM blanks should be used for data transmission and data archiving. Additionally, the IBM-compatible computer is the most universal desktop computing system at this time, so all CD-ROMs should be designed for this application.

Table 1 – Recommended Data Collection Approach and Archival Media and Media Formats

Data Types	Collection Media and Formats	Conversion Approach	Standard Media and Formats
Observations	Audio recordings – Cassette tapes	Voice recognition software	CD-ROM – TXT, RTF, DOC, or PDF files
	Written notes on paper	Human transcription to computer	CD-ROM – TXT, RTF, DOC, or PDF files
Questionnaires	Paper questionnaire forms	Human transcription to computer	CD-ROM – TXT, RTF, DOC, or PDF files
	Computerized questionnaires	Conversion software	CD-ROM – TXT, RTF, DOC, or PDF files
Surveys	Free-form written responses	Human transcription to computer	CD-ROM – TXT, RTF, DOC, or PDF files
	Audio recordings	Voice recognition software	CD-ROM – TXT, RTF, DOC, or PDF files
Images	Still photography with film	Film scanners	CD-ROM, using still picture formats such as Windows Bitmap (BMP), Windows Metafile (WMF), JPEG, or GIF files.
	Video photography on tape	Reflection copy scanners	
	Digital photography, with magnetic, CD-ROM, IR port	Computerized video capture	CD-ROM, using MPEG or QuickTime file formats
Sounds	Audio recorders, using magnetic tape	Computerized conversion	CD-ROM, using WAV, AU, or MP3 files
	Sound pressure meters using magnetic tape	Computerized conversion	CD-ROM, using WAV, AU, or MP3 files
Vibrations	Accelerometers using magnetic tape or telemetry	(Application specific)	ASCI, Lotus, EXCEL, dBase, or DIF files
Aircraft performance	A/C data buss, using magnetic tape or telemetry	(Application specific)	ASCI, Lotus, EXCEL, dBase, or DIF files

4.8.3 Data Analysis

Inferential and descriptive statistics express HPM data in terms of the population in a manner that encourages confidence in their accuracy and generalizability. Within the behavioral sciences, there is an implicit expectation that the relationship found or the effects produced will be probabilistic in nature. (For symmetrical normal data, the probabilities are normally distributed. Reaction time and time to repair, however, are typically in a Weibull distribution.) There is an equally explicit understanding if action is suggested or required. An inferential statistical test *should* demonstrate that the results or conclusions have less than a 5% probability of having occurred by chance (Orlady, *et al*., 1988, p. 44). There are currently no guidelines for interpreting results that do not meet this criterion. Appendix A of AFHRL's "OT&E Handbook for Aircrew Training Devices, Operational Effectiveness Analysis" provides guidelines for descriptive, inferential, and correlational statistical procedures. Additional guidelines are provided in Kesler, et al. (1981). Linnet and Brandt (1986) describe a bootstrap procedure to reduce positive bias in small samples (i.e., less than 25 subjects). Wasserman and Bockenholt (1989) describe how the bootstrap procedure can be used to analyze complex distributions of data.

4.8.4 Data Storage

In the absence of any overriding requirements (e.g., critical space limitations or equipment-specific formats), it is recommended that electronic data be stored as ASCII characters. Each data item should be delimited by a tab character, in accordance with the data entry requirements of virtually all spreadsheet and statistical analysis packages. This broad portability compensates for the significant penalty in storage efficiency. Due to the typically voluminous nature of physiological data, tab delineation is often inappropriate for this category of data.

For data files that contain a number of measures, it is recommended that the first record of the file consist of an ASCII descriptor file containing names of the measures in the fields that follow (as shown in Figure 4). This provides embedded documentation of the file format.

RECORD 1

Name of Meas 1	Tab	Name of Meas 2	Tab	Name of Meas 3	Name of Meas 4	Tab	(etc.)

RECORD 2

Value 1, Meas 1	Tab	Value 1, Meas 2	Tab	Value 1, Meas 3	Value 1, Meas 4	Tab	(etc.)

RECORD 3

Value 2, Meas 1	Tab	Value 2, Meas 2	Tab	Value 2, Meas 3	Value 2, Meas 4	Tab	(etc.)

RECORD 4

Value N, Meas 1	Tab	Value N, Meas 2	Tab	Value N, Meas 3	Value N, Meas 4	Tab	(etc.)

Note: "Meas" is short for "measure," the identification of the data item being collected (e.g., stick deflection).

Figure 4 – Suggested File Format

In choosing a database management system (DBMS) for human performance data, the researcher should consider the various commercially available DBMS products that comply with the Structured Query Language (SQL) requirements. SQL is an industry-standard DBMS access method that will promote the transportability of databases across different host computers and different DBMS packages.

4.8.5 Data Reporting

Reports of human performance data should include the definition of the behavior, dimensions and statistics, as discussed below.

4.8.5.1 Behavior

The type of activity should be specified, to include both discrete and continuous performance. Discrete performance includes actions such as switch actuation and voice commands; continuous behaviors include manual control and vigilance activities. The level of specificity should be stated. The manner in which tasks are described depends on the task taxonomy selected. Many task taxonomies have been developed (for example, Berson and Crooks, 1976; Christensen and Mills, 1967; Gagne, 1966). Mil-Hdbk-46855A cites a taxonomy that consists of functions, tasks, and steps or elements.

4.8.5.2 Dimensions

The dimensions of the human performance measurement should be stated, to include time, accuracy, quantity, and rate measures.

4.8.5.3 Statistics

The statistical aspects of the reported HPM data should be reported by giving the name of the statistic, the degrees of freedom, the numerical value, the mean squared error, and the probability level, as shown below:

$t(15) = 16.13$, mse = 0.591, $p < 0.01$

$F(2, 10) = 5.67$, mse = 0.123, $p < 0.05$

It is not necessary to cite references for well-known statistical procedures (e.g., t test, Mann-Whitney U, Dunnett's test, hierarchical ANOVA), but it is wise to cite references for new or unusual procedures. Complete ANOVA summary tables should be given only for complex analyses. For simple ANOVA, F ratios are given in the text. If the complete ANOVA table is given, the columns should be in this order: variable name, mean, degrees of freedom, and F ratio. In the table, asterisks should be used to indicate significant F values and should include a legend at the bottom of the table. The usual convention is *p< 0.05, ** p < 0.01; but any reasonable values less than 0.05 may be used.

The mean, standard deviation, and distribution should be included when reporting group performance data. An indication of the data variability should be included in data profiles based on plotted mean values.

5 Human Performance in the Context of Systems

5.1 Types of Testbeds

Testbeds for measurement include both physical location (environments) and devices. Measurement environments include the experimental laboratory, the engineering facility, the special test site (outside the engineering facility), and the operational environment (the so-called "real world"). Measurement devices include special laboratory instrumentation, special computer software, the mockup, the simulator, the prototype, and operational equipment.

The simulator reproduces the essential characteristics of operational equipment, although, as in the case of aircraft development, it may be built well before the operational equipment is produced. It performs like the operational equipment as far as personnel perception is concerned, but uses mechanisms, e.g., accelerators, that the operational equipment does have to produce operational effects.

The prototype is the first or one of the first operational equipment units produced and is used exclusively for test purposes. The operational equipment, which is the final configuration of its class, may also be used to test personnel proficiency.

All of these test environments and devices vary in terms of realism (similarity to operational equipment and situations) and the degree of control the investigator can exercise over his or her measurement. The relationship between realism and control is roughly inverse—the greatest control is found in the laboratory with the least realism, and the reverse is true of the operational equipment exercised in the operational environment. The mockup and the computer shell may be only crude approximations of the final system; the simulator may be much more realistic but more inflexible.

Before beginning a study, the researcher must decide how important control is vis-a-vis realism, and decide on the most suitable testbed. Often, he or she has no choice in the matter, the decision being made by considerations of development phase, cost, time, and managerial preference. Whatever the testbed, however, the investigator should realize in advance what limitations that testbed imposes on the study.

5.2 Human Performance Analysis and Synthesis

It is necessary to differentiate among individual performance measures (i.e., measure of an individual human's performance), group performance measures (measurement of team performance), and system performance measures. Each system also performs within the context of other systems and must be interoperable and compatible with these systems in a "system-of-systems" concept.

The US Air Force Operational Test and Evaluation Center (AFOTEC) uses a strategy-to-task analytic methodology to determine measures of effectiveness (MOE) and measures of performance (MOP). Strategy-to-task was developed by Rand Corp and presents a hierarchical method for linking national goals and interests with operational activities at the tactical engagement level. It links national goals and interests, to national security objectives, to national military objectives, to regional/campaign objectives, down to operational objectives and operational tasks. The tester then analyzes these tasks and determines the MOEs, which are defined as "a measure of a system's degree of accomplishment of a specific operational task." The MOEs are further deconstructed to MOPs which is "a qualitative or quantitative measure of a system's capabilities or characteristics. It indicates the degree to which that capability or characteristic performs or meets the requirement under specified conditions." For example, a new sonar system has a rated capability of sound ranging to X miles; this is a system measure of performance. In actual usage of the system, sonar personnel have been found to be able to detect targets at X-n miles with 63% correct target classifications and a 25% false-alarm rate. These last hypothetical, illustrative figures represent the performance of the personnel subsystem. By determining whether these MOPs meet the pre-established evaluation criteria, the overall MOE for the effectiveness of the sonar can be determined.

The MOEs of the individual systems can be aggregated to determine overall effectiveness of the system of systems.

6 Performance-Shaping Factors

6.1 Identification of Factors and Their Effect

Human performance is influenced by the single or combined effects of numerous factors that may enhance or degrade, facilitate, or interfere with that performance, depending upon the particular situation. Performance-shaping factors (PSFs) are aspects of the operator-machine system that can influence behavior and affect the time or accuracy of the human response (Swain and Guttman, 1980). To be relevant, a PSF must have the potential to significantly affect the magnitude of, frequency of, or variability in human performance. Identifying PSFs and establishing their effect on human performance are important aspects of HPM, for failure to account for PSFs may limit the validity and generalizability of HPM data.

A key consideration in dealing with PSFs is sensitivity. It is clear that a user could not account for all the PSFs that conceivably could influence HPMs in a specific task context. The user must identify those PSFs most likely to either enhance or degrade performance in that particular context. Sensitivity of performance to PSFs will vary, depending on the type of system, its general composition, and the environment within which it is operated. For example, vibration might be an important factor in measuring performance in an airborne surveillance context, but would have considerably less (if any) effect if the surveillance system were ground-based. Furthermore, there are complex interrelationships between and among stressors; for example, sleep loss actually reduces the size of the performance impairment found with noise (Hockey, 1983).

Another aspect of PSFs is long term or steady state performance of the subject. Cognitive and physical tasks can be assessed for peak performance and sustained performance. Often, sustained performance can be approximated from peak performance as is done with measurements from maximal weight lifting tests. Particularly concerning are ergonomic factors, given the spate of repetitive motion disorders related to work activities. Sample testing may fail to reveal problems with sustained performance or repetitive use unless the change in the condition of the subjects is assessed. Thus the man-machine system may change but the variable and its state is unknown.

Table 2 provides a general listing of PSFs, organized in five major categories (adapted from Blanchard (1973) and Kantowitz and Sorkin (1983)). These factors listed are by no means exhaustive but indicate the kinds of factors that can mediate human performance in various contexts. Four of the most widely influential PSFs (stress, workload, motivation, and training) are discussed in detail in the subsections that follow. A method for optimally characterizing the effects of PSFs is described in Metwally, et al. (1984). An alternative approach is described by Ringeisen and Shingledecker (1981). Lovesey (1990) has identified performance-shaping factors that should be considered during each step in system design.

6.2 Stress

Stress is a reaction to what is perceived to be a threat to either the individual's security or his/her accomplishment of the assigned task. Stress may also be a reaction to a physical condition, like excessive heat or cold, that challenges the body's protective mechanisms. Stressors arise from many sources: environmental conditions, situational uncertainty, risk of physical harm, heavy workload, or information overload.

While some amount of stress is recognized to have positive impacts on performance (as discussed under motivational PSFs), the effects of stress on performance are usually negative. There is evidence that, when placed under stress, the subject's focus of attention is more restricted, and fewer cues are sampled. The subject normally adapts to high stress by attending only to perceptual channels that are felt to be most important, based upon their costs, values, and salience to the assigned tasks (Wickens, 1984).

What factors should be considered in assessing sensitivity of the HPM to stress? Shaw and Riskind (1983) compared the frequency of stress-related illnesses with the dimensions of the jobs in which these illnesses occurred. The most highly correlated dimensions suggest the types of task situations in which stress may be particularly significant (note that these data are still open to the alternative explanation that people with health weaknesses migrate toward jobs with these characteristics):

(1) using various sources of information;

(2) making decisions;

(3) performing controlled manual or related activities, or both;

(4) communicating judgments or related information; and

(5) working in business-like settings.

Table 2 – Representative Performance-Shaping Factors

Operational Factors	Personnel Factors
1. System objectives 2. Doctrine; tactics; directives 3. Time on station 4. Time deployed 5. Quality control 6. Geographic location 7. Competition's objectives 8. Competition's tools	1. Training (on-the-job training, formal) 2. Team/individual experience 3. Motivation 4. Skill level (capacity) 5. Monotony and boredom 6. Fatigue 7. Clothing 8. Educational level 9. Morale/attitudes 10. Workload (physical and mental) 11. Anxiety and fear 12. Sensory deprivation 13. Sex; age; height; weight 14. Military/civilian/student 15. Stress 16. Arousal and vigilance 17. Circadian rhythm 18. Drugs and alcohol
Equipment Factors	
1. Physical parameters 2. Operating characteristics 3. Panel/console layout 4. Workspace layout 5. Reliability/maintainability 6. Operating mode 7. Operational status	
Task Factors	**External Environmental Factors**
1. Job aids 2. Complexity/difficulty 3. Task duration 4. Repetitiveness 5. Individual or team performance 6. Continuous/discrete 7. Supervision 8. Feedback 9. Task loading 10. Criticality stress 11. Work/rest cycle 12. Stimulus characteristics 13. Response characteristics 14. Task cueing	1. Temperature 2. Illumination 3. Distracting stimuli 4. Space limitations 5. Vibration 6. Vehicle motion 7. Gravitational/acceleration forces 8. Noise level 9. Turbulence 10. Visibility 11. Sea state 12. Wind velocity 13. Night/day 14. Ambient press

Stress has three components: (1) the perception that the subject has been or is being stressed; (2) physiological indices, such as rapid heartbeat or shortness of breath; and (3) performance effects of stress. Thus, there are three ways to measure stress: (1) subjective reports, (2) physiological measures, and (3) performance measures. While subjective reports of stress offer excellent face validity, they can fail to correlate with physiological or performance measures (e.g., Fenz and Epstein, 1967).

Using performance measures to gauge stress often intrudes on the primary objectives of HPM. There may be legal restrictions on putting subjects into very stressing situations when measuring performance in high-risk environments. The use of effective motivators in simulated environments (4.4) may induce levels of stress characteristic of the actual environment.

6.3 Workload

A multi-dimensional concept, workload is defined as the effort expended by the human operator in accomplishing the imposed task requirements (see Task 3). The task requirements (taskload) are the goal to be achieved, the time permitted to perform the task, and the performance level to which the task is completed. The factors affecting the effort expended are the information and equipment provided, the task environment, the subject's skills and experience, the strategies adopted, and the emotional response to the situation. This definition provides a testable link between taskload and workload. For example

(paraphrased from an example given by Azad Madni, Vice President, Perceptronics, Woodland Hills, CA, on the results of an Army study), the workload of a helicopter pilot in maintaining a constant hover may be 70 on a scale of 0 to 100. Given the task of maintaining a constant hover and targeting a tank may also be 70. The discrepancy results from the pilot self-imposing a strict performance requirement on hover-only (no horizontal or vertical movement), but relaxing the performance requirement on hover (a few feet movement) when targeting the tank to keep workload within a management level. These definitions allow taskload and workload to be explainable in real work situations. These definitions also allow logically correct analysis of taskload and workload. Realistically speaking, workload can *never* exceed 100% (a person cannot do the impossible). Any theories or reported results which allow workload to exceed 100% are not realistic. However, as defined, taskload may exceed 100%. An example is measuring "time required/time available." By the proposed definition, this taskload measurement may exceed 100% if the performance requirements are set too high (thereby increasing the time required) or the time available is set too low. The bottom line is workload cannot exceed 100%, even if taskload does exceed 100%.

Table 3 – Factors in Workload Assessment

Workload Aspect	Representative Questions	Candidate Measure	Comments
Sustained Workload	• What was the average overall workload? • How will various intensities of sustained workload affect performance?	• Subjective rating scales • Overall performance on salient tasks	
Momentary Workload	• What was the magnitude of workload during peak periods? • How was human performance affected during periods of high demand?	• Subjective rating scales • Secondary task performance	• Global measures of task performance are inappropriate
Reserve Capacity	• What margin of full performance did this task require? • How much more can the operator handle effectively?	• Secondary tasks • Subjective rating scales	• Primary task measures are inappropriate

6.4 Situation(al) Awareness

Situation awareness, also termed situational awareness or SA, is an operator's internal mental model of the surrounding world. The quality of this internal model is critical for effective decision making. Systems operators must ascertain the current status and dynamics of their systems, and the status and dynamics of other relevant elements in the environment, in order to determine the best course of action to take at any point in time. Without this knowledge, most operators will not be able to function satisfactorily. This is true for many different types of systems, including aircraft; air traffic control; large systems, such as flexible manufacturing systems, refineries, and nuclear power plants; strategic systems such as fire fighting units, certain police units and military command centers; and for many daily activities, such as driving. As decision selection and performance flow directly from this situational understanding, often in an automatic fashion, the formation of correct and complete SA is critical.

Like workload, it may be desirable to ascertain the level and quality of an operator's situation awareness directly as a primary measure of interest. This is because, in testing situations, large variabilities in direct performance measures can mask true differences in workload or SA that can be significant in system usage. Thus, the direct measurement of these PSFs can provide highly useful knowledge for system design and training efforts.

6.5 Motivation

A subject's performance can vary dramatically with the presence of motivating factors. Incentives such as augmented feedback, financial reward, or induced competitiveness have led to net improvements in speed and accuracy, alertness (arousal level), selectivity (ability to prioritize concurrent tasks), and short-term memory (Hockey, 1983).

Although there are no reliable and valid means of measuring the subject's level of motivation, there are ways to elevate it. For example, when pilots participate in ground-based simulation of air combat, a scoring strategy is sometimes used as a motivational technique. Based upon realistic mission objectives, the scores provide an easy and immediate way for pilots to compete individually and as teams. In actuality, the scores frequently have no direct contribution to the test objectives or results, but they foster the competitive and highly motivated aspects of actual combat (Lehman *et al.*, 1989).

In deriving motivating factors, it is important to strive for a level of motivation characteristic of the normal task environment. It is also important that the motivators correspond to the objectives of the task for which performance is being measured. For example, if the objective is predominantly one of survival (defense), the motivators should not give undue weight to offensive actions.

6.6 General State of Health

Testing is often limited in the number of subjects available or affordable. Variables include wellness, current or recent illness or injury, and fatigue. Chronic illness or treatment for the same may pose limitations. Certain blood pressure medicines constrict the pupil size and limit performance under low light conditions. Recent illness may cause fatigue or asthenia or diminish cognitive capability. Recent injury may limit range of motion or strength. Acute or chronic fatigue can also diminish physical or cognitive performance. Such subjects may be identified as outliers or may skew data for a small sample population.

6.7 Physiological Capacity

The proposed task may exceed the capabilities of the subject population or potential users. Variables include visual activity, aural acuity, and anthropometric data. Hearing or vision deficits may require corrective aids that are not used for reasons of unawareness of the deficit or vanity. Height or weight may limit the space in which a person can function, such as coach seating on an aircraft.

6.8 Training

Training is a factor that can significantly affect the performance measurement. If performance is plotted against number of trials, the resulting J curve indicates that substantial improvement occurs in the early trials and very little improvement occurs in later trials. The point at which there is little continued improvement is called asymptotic learning. Unless subjects are trained to asymptote before the first trials, their performance data will be skewed by the learning process, regardless of the experimental conditions.

Training is particularly important if human performance is being used as a means of evaluating the acceptability of a human-machine system, because poorly trained subjects can make a system look worse than it is. Equipment, no matter how appropriate for the given task, cannot be adequately evaluated unless test subjects are suitably trained. Knowledge of the equipment and tasks must be assessed prior to testing. Pre-test assessment of training assures that the subjects are sufficiently prepared and trained for the test. Additionally, the pre-test training assessment data, in conjunction with the test data, can be valuable in estimating training requirements for the expected users.

Training can be measured experimentally or against a standard. If we know, for example, what a pilot must be able to do in flight, an examiner can test the subject in the cockpit by observing his or her maneuvers and determining if he or she satisfies the standard. We do not test the training directly; we test the effects of training on performance.

Often (not always) what we want to know is how to determine whether a particular training method, device, or regime has been effective in modifying performance. The Transfer Effectiveness Ratio provides a quantitative means of measuring the amount of learning that is transferred from a training-device environment to an operational environment.

The transfer-of-training paradigm calls for a control group that has received either no training or some type of training other than the one in which the evaluator is interested. That other training may be given in another training device or as OJT.

Evaluating the effectiveness of training devices is assisted by a number of transfer-of-training experimental designs. These include: (1) interruption of pre/post-test design, (2) pre-existing transfer design, (3) device-to-device transfer design, (4) backward transfer design, (5) uncontrolled transfer design, and (6) device performance improvement design. All of these are discussed in Meister (1985).

7 Referenced Publications

The following publications are essential for effective use of this Guide:

AFHRL. *OT&E handbook for aircrew training devices, operational effectiveness analysis*. Williams Air Force Base, AZ; 1991.

American Psychological Association. *Publication manual*. Washington, DC; 1983.

Anastasi, A. *Psychological Testing (6th edition)*. New York: Macmillan; 1988.

Babbitt, B.A. and Nystrom, C.O. *Questionnaire construction manual (ARLI-RP-89-20)*. Fort Hood, TX: U.S. Army Research Institute; June 1989.

Bainbridge, L. *Verbal reports as evidence of the process operator's knowledge*. International Journal of Man-Machine Studies. 11, 411-436; 1979.

Baker, D.P. and Salas, E. *Principles for measuring teamwork skills. Human Factors*. 34:469-475; 1992.

Berson, B. T. and Crooks, W. H. *Guide for obtaining and analyzing human performance data in a material development project*. Aberdeen Proving Ground, MD: Army Human Engineering Laboratory; September 1976.

Biferno, M.A., Dennison, T.W. and Gawron, V.J. *Mockups, Physical and Electronic Human Models, and Simulations.* In Handbook of Human Factors Testing and Evaluation. Muhaw, New Jersey: Lawrence Erlbaum Associates; 1996.

Blanchard, R. E. *Requirements, concept, and specification for a Navy human performance data store*. New London, CT: Naval Underwater Systems Center, April 1973.

Brinkman, J.A. *Verbal protocol accuracy in fault diagnosis*. Ergonomics. 36(11), 1381-1397; 1993.

Christensen, J.M. and Mills, R.G. *What does the operator do in complex systems*. Human Factors. 4, 385-392; 1967.

Cordes, R.E. *The relationship between post-task and continuous-vicarious ratings of difficulty*. International Journal of Human-Computer Interaction. 2: 1993.

Cordes, R.E. *Are software usability tests biased in favor of your product?* Proceedings of the 28th Annual Technical Symposium of ACM; 1989a.

Cordes, R.E. *The "I know it can be done or you wouldn't have asked me" bias in usability evaluations*. Proceedings of the Human Factors/Product Usability ITL; 1989b.

Corwin, W. H., Sandry-Garza, D. L., Biferno, M. H., and Boucek, G. P., Jr. *Assessment of crew workload measurement methods, techniques and procedures. Volume II - Guidelines for the use of workload assessment techniques in aircraft certification*. Wright-Patterson Air Force Base, OH: Wright Research and Development Center; September 1989.

Crites, D.C. *Using the video tape method*. In Air Force Systems Command Design Handbook DH 1-3, Part 2, Series 1-0, General Human Factors Engineering, Chapter 7, Section DN 7E3, pp. 1-6. Washington, DC: U.S. Government Printing Office; 1980.

Cronbach, L.J. *Essentials of psychological testing*. New York: Harper & Row; 1970.

Crutcher, R.J. *Telling what we know: The use of verbal report methodologies in psychological research*. Psychological Science. 5(5): 241-244; 1994.

Defense Documentation Center. *Performance measurement: A DDC bibliography*. Alexandria, VA: Defense Documentation Center; September 1972.

Defense Science Board Task Force on Improving Test and Evaluation Effectiveness. Report. Washington, DC: Office of the Under Secretary of Defense for Acquisition; December 1989.

DoD-Hdbk-763. Department of Defense. *Human engineering procedures guide*. Washington, DC; 1987.

Edwards, M. R. and Verdini, W. A. *Engineering and technical management: accurate human performance measures = productivity*. Society of Research Adminis- trators Journal. 18(2): 5-19; Fall 1986.

Eggemeier, F.T., Biers, D.W., Wickens, C.D., Andre, A.D., Vreuls, D., Billman, E.R., and Schuerren, J. *Performance and workload analysis system: Analysis of candidate measures*. Wright-Patterson Air Force Base, OH: Air Force Human Systems Division; 1989.

Ericsson, K.A. and Simon, H.A. *Protocol analysis: Verbal reports as data*. Cambridge, MA:MIT Press; 1993.

Fenz, W.D. and Epstein, S. *Gradients of physiological arousal in parachutists as a function of an approaching jump*. Psychosomatic Medicine. 29(1):33-51; 1967.

Finley, D. L., Muckler, F .A., Gainer, C. A., and Obermayer, R. W. *An analysis and evaluation methodology for command and control*. Arlington, VA: Office of Naval Research; November 1975.

Gagne, R.M. (Ed.). *Psychological principles in system development*. New York: Holt, Rinehart and Winston; 1966.

Gawron, V.J. *How to design an experiment*. Buffalo, NY: Calspan Learjet Technical Memorandum Number 44; August 1989.

Gilhooly, K.J. and Green, C. *A suite of computer programs for use in verbal protocol analysis*. Literary and Linguistic Computing. 4(1): 1-5; 1989.

Green, A. *Verbal protocol analysis*. The Psychologist. March: 126-129; 1995.

Harker, S. *The use of prototyping in the development of large scale applications*. The Computer Journal. 31: 420-425; 1988.

Harris, R.L., Glover, B.J., and Spady, A.A. *Analytical techniques of pilot scanning behavior and their application* (NASA Technical Paper 2525). Hampton, VA: NASA/Langley; July 1986.

Hennessy, R. T. *Practical human performance testing and evaluation. In H.K. Boohrer (Ed.) MANPRINT: An approach to systems integration* (pp. 433-470). New York: Van Nostrand Reinheld; 1990.

Highhouse, S. *A verbal protocol analysis of choice under ambiguity*. Journal of Economic Psychology. 15:621-625, 1994.

Hockey, R. *The cognitive patterning of stress states*. In R. Hockey (Ed.) *Stress and fatigue in human performance*. New York, NY: Wiley and Sons; 1983.

Hutson, W. *Personal communication on dissertation to R. Cordes*; 1989.

Kantowitz, B.H. *Selecting measures for human factors research*. Human Factors. 34:387-398; 1992.

Kantowitz, B. H. and Sorkin, R. D. *Human factors: understanding people-system relationships*. New York: Wiley; 1983.

Kesler, G.P., Miller, J.J., and Goldner, R.R. *Data analysis handbook for ASW performance measurement*. Washington, DC: ASW Systems Project Office; December 1981.

Kliem, R. J. *Performance measurement: a necessary tool*. Computerworld. 16(32): 49; 9 August 1982.

Knowles, W. B., Burger, W. J., Mitchell, M.B., Hanifan, D. T., and Wulfeck, J. W. *Models, measures, and judgments in system design*. Human Factors. 11(6): 577-590; 1969.

Kramer, A. *Cortical evoked potentials*. In A. Gale and B. Christie (Eds.) *Psychophysiology and the electronic workplace*. Sussex, England: Wiley; 1987.

Kramer, A., Sirevaag, E.J., and Braune, R. *A psychophysiological assessment of operator workload during simulated flight missions*. In A. Gale and B. Christie (Eds.) *Psychophysiology and the electronic workplace*. Sussex, England: Wiley; 1987.

Lane, N.E. *Issues in performance measurement for military aviation with applications (NTSC-TR-86-008).* Orlando, FL: Naval Training Systems Center; April 1986.

Lehman, E. F., Jenkins, M.L., Chaffee, A., Dinkel, R., Dvorchak, S. R., Hamilton, W. L., Holmes, G., Masters, R. M., Pero, K., Reed, J. D., Schneider, R. R., Sparks, B. R., and Stuart, R. B. *A handbook for conducting pilot-in-the-loop simulation for crew station evaluation (HSD-TR-90-007)*. Wright Patterson Air Force Base, OH: Armstrong Aerospace Medical Research Laboratory; 1989.

Lehman, E. F., Gawron, V. J., Nelson, D. R., Cody, W. J., and Jenkins, M. L. *Handbook for crewstation test program planning (MSD-TR-91-0001)*. Wright Patterson Air Force Base, OH: Armstrong Aerospace Medical Research Laboratory; 1991.

Linnet, K. and Brandt, E. *Assessing diagnostic tests once an optimal cutoff point has been selected*. Clinical Chemistry. 32(7):1341-1346; 1986.

Lovesey, E.J. *Quantification of some of the factors that determine operational effectiveness*. Farnborough, UK: Royal Aerospace Establishment; 1990.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner, A.C., and Wherry, R.J. *Operator workload: comprehensive review and evaluation of operator workload methodologies (Technical Report 851).* Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences; June 1989.

Martin, I. and Venables, P.H. (Eds) *Techniques in psychophysiology*. New York: Wiley; 1980.

McGraw Hill Dictionary of scientific and technical terms. New York: McGraw-Hill; 1974.

Meister, D. *Behavior analysis and measurement methods*. New York: Wiley; 1985.

Meister, D. *Human factors testing and evaluation*. New York: Elsevier; 1986.

Meister, D. *Divergent Viewpoints: Essays on human factors questions*. 1995.

Metwally, A.M., Mohammed, F.A., and Omar, A.A. *Quantification of human performance in nuclear tasks*. Nuclear Power Performance and Safety Proceedings. AEA-CU-48/224; 71-77; 1984.

Mil-Hdbk-46855A. *Human engineering program process and procedures.* Washington D.C.: 1999.

Mil-Std-1472D. *Human engineering design criteria for military systems, equipment and facilities*. Redstone Arsenal, AL: US Army Missile Command; 14 March 1989.

Murphy, K.R. and Davidshofer, C.O. *Psychological testing: Principles and Applications*. Englewood Cliffs, NJ: Prentice Hall; 1991.

Obata, T., Daimon, T., and Kawashima, H. *A cognitive study of in-vehicle navigation systems applying verbal protocol analysis to usability evaluation*. IEEE Vehicle Navigation and Information Systems Conference Proceedings. 232-237; 1993.

Orlady, H. W., Hennessy, R. T., Obermayer, R. W., Vreuls, D., and Murphy, M. R. *Using mission simulation for human factors research in air transport operations*. Washington, DC: NASA TM 88330; 1988.

Orne, M.T. *On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications*. American Psychologist. 17:776-783; 1962.

Payne, J.W. *Thinking aloud: insights into information processing*. Psychological Science. 5(5): 241-247; 1994.

Ringeisen, R. D. and Shingledecker, C. A. *Combined stress and human performance: a weighted digraph model*. Mathematical Social Sciences. 1: 297-305; 1981.

Robson, J.I. and Crellin, J.M. *The role of user's perceived control in interface design, employing verbal protocol analysis*. Applied Ergonomics. 20(4):246-251; 1989.

Rosenthal, R. *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts; 1966.

Rosenthal, R. *Covert communication in the psychological experiment*. Psychological Bulletin. 67:356-367; 1967.

Shaw, B.J. and Riskind, J.H. *Predicting job stress using data from the positional analysis questionnaire*. Journal of Applied Psychology. 68: 253-262; 1983.

Smith, S.L. and Mosier, J.M. *Guidelines for designing user interface software* (Report ESD-TR-86-278). Hanscom AFB, MA: Electronic Systems Division, USAF Systems Command; 1986.

Svenson, O. *Scaling evaluative statements in verbal protocols from decision processes*. In P. Humphreys, O. Svenson, and A. Vari (Eds.) *Analyzing and aiding decision processes*. Amsterdam: North-Holland; 1983.

Stanley, D.H. and Campbell, J.C. *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally; 1966.

Swain, A.D., and Guttman, H.E. *Handbook of human reliability analysis and emphasis on nuclear power plant applications (NUREG/CR-1278).* U.S. Nuclear Regulatory Commission. Albuquerque, NM: Sandia Laboratories; 1980.

Swink, J. R., Butler, E. A., Lankford, H. E., Miller, R. M., and Watkins, H. *Definition of requirements for a performance measurement system for C-5 aircrew members*. Williams Air Force Base, AZ: Air Force Human Resources Laboratory; October 1978.

Van Cott, H.P. and Chapanis, A. *Human engineering tests and evaluation*. In H.P. Van Cott and R.G. Kinkade (Eds.) *Human engineering guide to equipment design*. Washington, DC: U.S. Government Printing Office; 1972.

Vineberg, R. and Joyner, J. N. *Performance measurement in the military services*. In F. Landy, S. Zedeck, J. Cleveland, and A. Landy (Eds.) *Performance measurement and theory*. Hillsdale, NJ: Lawrence Erlbaum; 1983.

Virzi, R.A. *Refining the test phase of usability evaluation: How many subjects is enough?* Human Factors. 34:457-468; 1992.

Vreuls, D. and Obermayer, R. W. *Human-system performance measurement in training simulators*. Human Factors. 27(3): 241-250; 1985.

Wasserman, S. and Bockenholt, U. *Bootstrapping: applications to psychophysiology*. Psychophysiology. 26(2): 208-221; 1989.

Wickens, C. D. *Engineering psychology and human performance*. Columbus, OH: Meville Publishing Company; 1984.

Williges, B.H. and Williges, R.C. *Dialogue design considerations for interactive computer systems*. In F.A. Muckler (Ed.) *Human factors review* 1984. Santa Monica, CA: The Human Factors Society, 167-208; 1984.

Williges, R. and Mills, R. *Catalog of methodological considerations for systems experimentation*. Wright-Patterson Air Force Base, OH: Air Force Armstrong Aerospace Medical Research Laboratory; June 1988.

Williges, R.C., Williges, B.H., and Han, S.H. *Developing quantitative guidelines using integrated data from sequential experiments*. Human Factors. 34: 399-408; 1992.

Wilson, T.D. *The proper protocol: Validity and completeness of verbal reports*. Psychological Science. 5(5): 249-252; 1994.

Annex A - Taxonomy of Human Performance Measures

This appendix is not part of this Guide but is included for information only.

A.1 Criteria

Two criteria were used to develop the task portion of the taxonomy presented in this appendix:

(1) the taxonomy must describe the dimensions of the task and its environment, and

(2) the taxonomy must describe what the task is, rather than how the operator performs the task (Berry, 1980).

Nine additional criteria (Companion and Corso, 1977, pp. 359-360) were used during the development of the complete taxonomy. Each measure:

(1) must simplify the description of tasks in the system,

(2) should be generalizable,

(3) must be compatible with other terms in general use,

(4) must be complete and internally consistent,

(5) must be compatible with the theory or system to which it will be applied,

(6) should help to predict operator performance,

(7) must have some utility,

(8) must be cost-effective,

(9) must provide a framework around which all relevant data can be integrated.

The taxonomy of HPM is presented in Table A1. A taxonomy of driver errors is presented in Table A2.

Table A.1 – Taxonomy of Performance Measures

1. Time (from Meister, 1985)
    - 1.1 Reaction time, i.e., time to:
        - 1.1.1 Perceive event
        - 1.1.2 Initiate movement
        - 1.1.3 Initiate correction
        - 1.1.4 Initiate activity following completion of prior activity
        - 1.1.5 Detect trend of multiple related events
    - 1.2 Time to complete an activity already in process, i.e., time to:
        - 1.2.1 Identify stimulus (discrimination time)
        - 1.2.2 Complete message, decision, control adjustment
        - 1.2.3 Reach criterion value
    - 1.3 Overall time (duration):
        - 1.3.1 Time spent in activity
        - 1.3.2 Percent time on target
    - 1.4 Time sharing among events
    - 1.5 Error characteristics:
        - 1.5.1 Amplitude measures
        - 1.5.2 Frequency measures
        - 1.5.3 Content analysis
        - 1.5.4 Change over time
2. Accuracy
    - 2.1 Corrections in observations, i.e., accuracy in:
        - 2.1.1 Identifying stimuli internal to system
        - 2.1.2 Identifying stimuli external to system
        - 2.1.3 Estimating distance, direction, speed, time
        - 2.1.4 Detecting stimulus change over time
        - 2.1.5 Detecting trend based on multiple related events
        - 2.1.6 Recognizing signal in noise
        - 2.1.7 Recognizing out-of-tolerance condition
    - 2.2 Response-output correctness, i.e., accuracy in:
        - 2.2.1 Control positioning or tool usage
        - 2.2.2 Reading displays
        - 2.2.3 Symbol usage, decision-making, and computing
        - 2.2.4 Response selection among alternatives
        - 2.2.5 Serial response
        - 2.2.6 Tracking
        - 2.2.7 Communicating
3. Amount achieved or accomplished
    - 3.1 Degree of success
    - 3.2 Percentage of activities accomplished
    - 3.3 Measures of achieved reliability (numerical reliability estimates).
4. Frequency of occurrence
    - 4.1 Number of responses per unit, activity, or interval:
        - 4.1.1 Control and manipulation responses
        - 4.1.2 Communications
        - 4.1.3 Personnel interactions
        - 4.1.4 Diagnostic check
    - 4.2 Number of performance consequences per activity, unit, or interval:
        - 4.2.1 Number of errors
        - 4.2.2 Number of out-of-tolerance conditions
    - 4.3 Number of observing or data-gathering responses:
        - 4.3.1 Observations
        - 4.3.2 Verbal or written reports
        - 4.3.3 Requests for information
        - 4.3.4 Rate of engagement

5. Behavior categorization by observers
    5.1 Judgment of performance:
        5.1.1 Rating of operator/crew performance adequacy
        5.1.2 Rating of task or mission segment performance adequacy
        5.1.3 Estimation of amount (degree) of behavior displayed
        5.1.4 Measures of achieved maintainability
        5.1.5 Equipment failure rate (mean time between failures)
        5.1.6 Cumulative response output
        5.1.7 Proficiency test scores (written)
    5.2 Magnitude achieved:
        5.2.1 Terminal or steady-state value (e.g., temperature high point)
        5.2.2 Changing value or rate (e.g., degree changes per hour)
6. Consumption or quantity used
    6.1 Resources consumed per activity:
        6.1.1 Fuel/energy conservation
        6.1.2 Units consumed in activity accomplishment
    6.2 Resources consumed by time
    6.3 Subjective reports:
        6.3.1 Interview content analysis
        6.3.2 Self-reporting of experience (“debriefing”)
        6.3.3 Peer, self, or supervisor ratings
        6.3.4 Analysis of operator/crew behavior characteristics
        6.3.5 Determination of behavior relevance:
            6.3.5.1 Omission of relevant behavior
            6.3.5.2 Occurrence of nonrelevant behavior
        6.3.6 Casual description of out-of-tolerance condition
7. Workload:
    7.1 Subjective
    7.2 Performance
8. Probability:
    8.1 Probability of kill (PK)
    8.2 Probability of survival (P_S)
    8.3 Probability of completing task
    8.4 Probability of error
    8.5 Likelihood ratio (λ)
    8.6 Probability of hit (P_H)
9. Space/Distance (e.g., CEP)
10. Errors (from Boohrer, 1990, pages 244-245)
    10.1 Observation of system state:
        10.1.1 Excessive - improper rechecking of correct readings of appropriate state variables
        10.1.2 Misinterpreted - erroneous interpretation of correct readings of appropriate state variables
        10.1.3 Incorrect - incorrect readings of appropriate state variables
        10.1.4 Incomplete - failure to observe sufficient number of appropriate state variables
        10.1.5 Inappropriate - observations of inappropriate state variables
        10.1.6 Lack - failure to observe any state variables
    10.2 Choice of hypothesis:
        10.2.1 Inconsistent - could not cause particular values of state variables observed
        10.2.2 Unlikely - could cause values observed, but much more likely causes should be considered first
        10.2.3 Costly - could cause values observed, but very costly (in time or money)
        10.2.4 Irrelevant - does not functionally relate to state variables observed
    10.3 Testing of hypothesis:
        10.3.1 Incomplete - stopped before reaching a conclusion
        10.3.2 Acceptance - reached erroneous conclusion

- 10.3.3 Rejection - considered and discarded correct conclusion
- 10.3.4 Lack - hypothesis not tested

10.4 Choice of goal:

- 10.4.1 Incomplete - insufficient specification of goal
- 10.4.2 Incorrect - choice of counterproductive goal
- 10.4.3 Unnecessary - choice of nonproductive goal
- 10.4.4 Lack - goal not chosen

10.5 Choice of procedure:

- 10.5.1 Incomplete - choice would not fully achieve goal
- 10.5.2 Incorrect - choice would achieve incorrect goal
- 10.5.3 Unnecessary - choice unnecessary for achieving goal
- 10.5.4 Lack - procedure not chosen

10.6 Execution of procedure:

- 10.6.1 Omitted - required step omitted
- 10.6.2 Repeated - unnecessary repetition of required step
- 10.6.3 Added - unnecessary step added
- 10.6.4 Sequence - required steps executed in wrong order
- 10.6.5 Timing - step executed too early or too late
- 10.6.6 Discrete - discrete control in wrong position
- 10.6.7 Continuous - continuous control in unacceptable range
- 10.6.8 Incomplete - stopped before procedure complete
- 10.6.9 Unrelated - unrelated inappropriate step executed
- 10.6.10 Incorrect grasping — wrong contact with objects (Dubrovsky, 1985)
- 10.6.11 Failure to \\\check results — failure to compare the goal and the outcome (Dubrovsky, 1985)
- 10.6.12 Disapproaching — “incorrect departure (unintended activating of a control during departure from the just used control)” (Dubrovsky, 1985, p. 905).

Table A.2 – Taxonomy of Driver Errors

1. Error
    - 1.1 Not noticing pedestrians crossing
    - 1.2 Not noticing 'Give Way' sign
    - 1.3 Misjudge an overtaking gap
    - 1.4 Failing to notice a cyclist
    - 1.5 Overtake a right turner
    - 1.6 Not checking mirror
    - 1.7 Brake too quickly on a slippery road
    - 1.8 Insufficient attention to vehicle ahead
2. Violation
    - 2.1 Unofficial races
    - 2.2 Close following
    - 2.3 Overtaking on the inside
    - 2.4 Breaking speed limit late/early in day
    - 2.5 Giving chase
    - 2.6 Not stopping at red lights
    - 2.7 Showing hostility to other drivers
    - 2.8 Drinking and driving
3. Lapse
    - 3.1 Hitting something reversing
    - 3.2 Wrong lane at roundabout/junction
    - 3.3 Forget where the car is in a car park
    - 3.4 No recollection of road travelled
    - 3.5 Take more usual route by error
    - 3.6 Go for the wrong switch
    - 3.7 Take wrong exit from the roundabout (Parker, Reason, Manstead, and Stradling, 1995, p. 1042)

A.2 References

Berry, G.L. *Task taxonomies and modeling for system performance prediction*. Proceedings of the Human Factors Society 24th Annual Meeting, 425-429; 1980.

Boohrer, H. (Ed.) *MANPRINT: An approach to systems integration*. New York: Van Nostrand Reinheld; 1990.

Companion, M. A. and Corso, G. M. *Task taxonomy: Two Ignored Issues*. Proceedings of the Human Factors Society 21st Annual Meeting, 358-361; 1977.

Dubrovsky, V. *A taxonomy of human errors based upon the structure of an action*. Proceedings of the International Conference on Cybernetics and Society, 903-907; 1985.

Meister, D. *Behavior analysis and measurement methods*. New York: Wiley; 1985.

Parker, D., Reason, J.T., Manstead, A.S.R., and Stradling, S.G. *Driving Errors, Driving Violations, and Accident Involvement*. Ergonomics. 38(5), 1036-1048, 1995.

American Institute of Aeronautics and Astronautics

1801 Alexander Bell Drive, Suite 500
Reston, VA 20191-4344

ISBN 1-56347-451-4