

Guide

Guide to Human Performance Measurements

AIAA standards are copyrighted by the American Institute of Aeronautics and Astronautics (AIAA), 1801 Alexander Bell Drive, Reston, VA 20191-4344 USA. All rights reserved.

AIAA grants you a license as follows: The right to download an electronic file of this AIAA standard for temporary storage on one computer for purposes of viewing, and/or printing one copy of the AIAA standard for individual use. Neither the electronic file nor the hard copy print may be reproduced in any way. In addition, the electronic file may not be distributed elsewhere over computer networks or otherwise. The hard copy print may only be distributed to other employees for their internal use within your organization.

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

Copyright © 2001 American Institute of Aeronautics and Astronautics
All rights reserved.

No part of this publication may be reproduced in any form, in an electronic retrieval system or
otherwise, without prior written permission of the publisher.

Printed in the United States of America

Contents

Foreword	vi
1 Scope and Purpose.....	1
1.1 Scope.....	1
1.2 Purpose.....	1
2 Vocabulary.....	1
3 Applications	3
3.1 General Applications	3
3.2 Basic Research	5
3.3 Applied Research	6
3.4 System Design and Development.....	7
3.4.1 Scope of this Application.....	7
3.4.2 The Checklist.....	7
3.4.3 Mockup Testing	9
3.4.4 Rapid Prototyping.....	9
3.4.5 Verbal Protocol Methodology.....	11
3.5 System Test and Evaluation.....	12
4 General Issues in Human Performance Measurement.....	15
4.1 Underlying Problems in Human Performance Measurement	15
4.2 Bridging the Gap Between Human and System Performance Measurement.....	16
4.2.1 Interactions of Human and System Performance and Their Exploration	16
4.2.2 Developing Performance Criteria	16
4.2.3 Using Expert Judgments of Task Performance.....	17
4.2.4 Modeling and Simulation.....	17
4.3 Asking the Right Questions	17
4.4 Selection Criteria for Human Performance Measures.....	18
4.4.1 Appropriate Level of Detail	19
4.4.2 Reliability.....	20
4.4.3 Validity	21
4.4.4 Sensitivity.....	21
4.4.5 Diagnosticity	22
4.4.6 Non-intrusiveness.....	22
4.4.7 Implementation Requirements.....	23
4.4.8 Operator Acceptance	23
4.4.9 Fairness	23
4.4.10 Accuracy	23

4.4.11	Simplicity	24
4.4.12	Timeliness	24
4.4.13	Objectivity	24
4.4.14	Quantitativeness/Qualitativeness	24
4.4.15	Cost	24
4.4.16	Flexibility	24
4.4.17	Utility	24
4.5	Measurement Uncertainty	25
4.5.1	Combining Errors	25
4.5.2	Outliers	25
4.5.3	Speed/Accuracy Tradeoff	25
4.5.4	Biases in Usability Measurement	25
4.6	Experimental Design	28
4.7	Instrumentation	28
4.7.1	Accelerometer	29
4.7.2	Anemometer	30
4.7.3	Anthropometry Instrument Kit	30
4.7.4	Electrogoniometer	30
4.7.5	Force, Torque, and Dimension Kit	30
4.7.6	Gas Tester	30
4.7.7	Hygrometer or Psychrometer	30
4.7.8	Motion Measurement Systems	30
4.7.9	Photometer	31
4.7.10	Sound Level Meter and Analyzer	31
4.7.11	Spot Brightness Meter	31
4.7.12	Thermometer	31
4.7.13	Vibration Meter and Analyzer	31
4.7.14	Video Tape Instrumentation	31
4.7.15	Spectroradiometer	32
4.7.16	Video Digitizer	32
4.7.17	Digital Audio Tape (DAT) Recorder	32
4.8	Data Collection and Analysis	32
4.8.1	General Considerations	32
4.8.2	Data Collection	32
4.8.3	Data Analysis	35
4.8.4	Data Storage	35

4.8.5	Data Reporting	36
5	Human Performance in the Context of Systems	36
5.1	Types of Testbeds.....	36
5.2	Human Performance Analysis and Synthesis	37
6	Performance-Shaping Factors.....	37
6.1	Identification of Factors and Their Effect	37
6.2	Stress	38
6.3	Workload.....	39
6.4	Situation(al) Awareness	40
6.5	Motivation	41
6.6	General State of Health.....	41
6.7	Physiological Capacity	41
6.8	Training	41
7	Referenced Publications	42
	Annex A - Taxonomy of Human Performance Measures	47
A.1	Criteria.....	47
A.2	References	51

List of Figures

Figure 1 – Major Steps in Measurement.....	4
Figure 2 – The Checklist Comparison Process	8
Figure 3 – Types of HPM Instrumentation	29
Figure 4 – Suggested File Format.....	35

List of Tables

Table 1 – Recommended Data Collection Approach and Archival Media and Media Formats.....	34
Table 2 – Representative Performance-Shaping Factors.....	39
Table 3 – Factors in Workload Assessment.....	40
Table A1 – Taxonomy of Performance Measures	48
Table A2 – Taxonomy of Driver Errors.....	51

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.

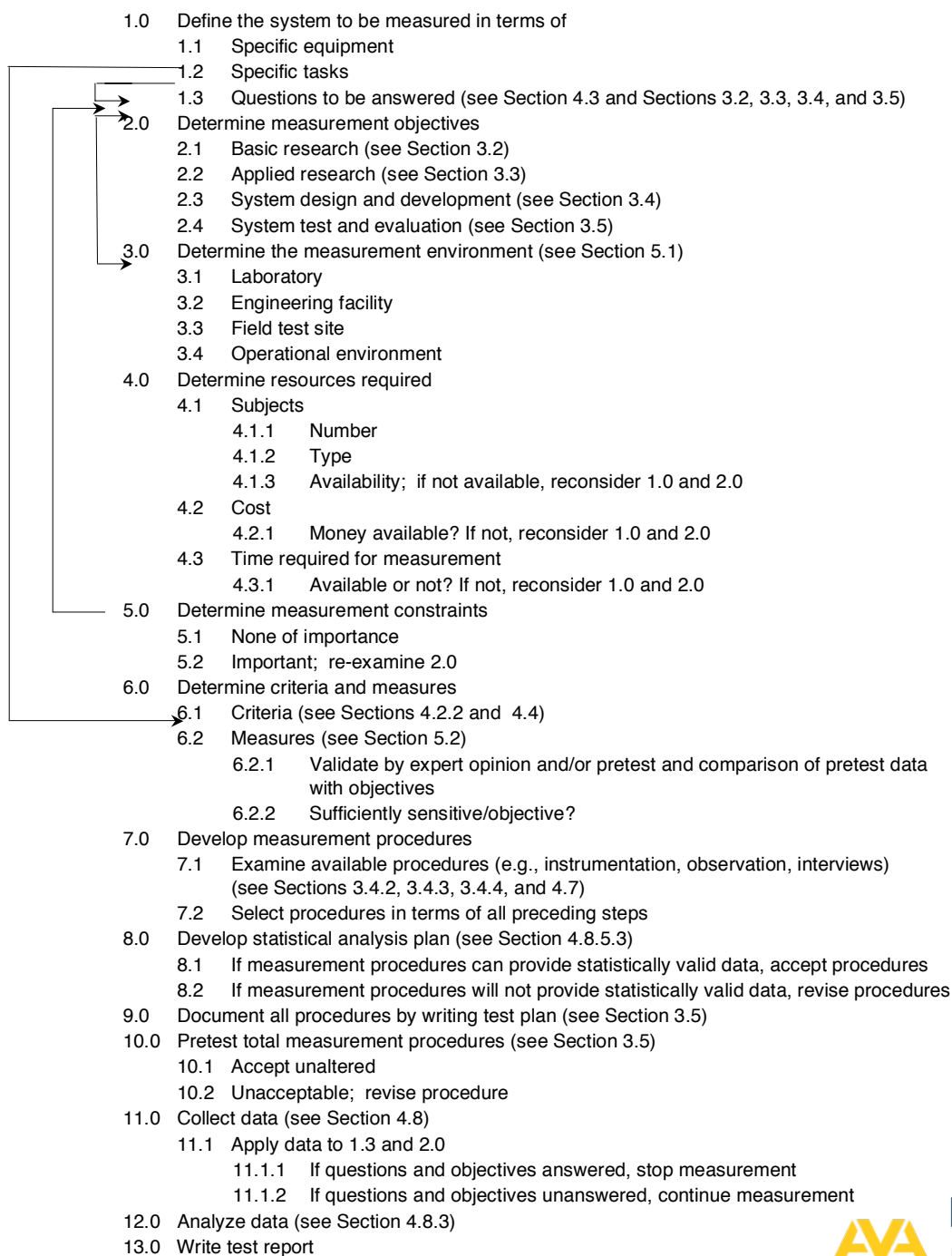


Figure 1 – Major Steps in Measurement

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 latter's 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.

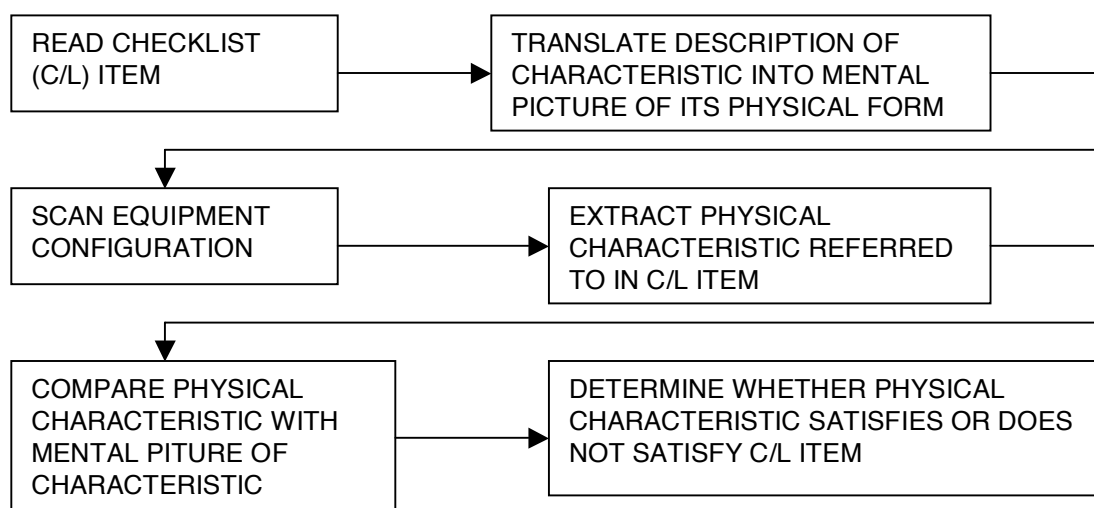


Figure 2 – The Checklist Comparison Process

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