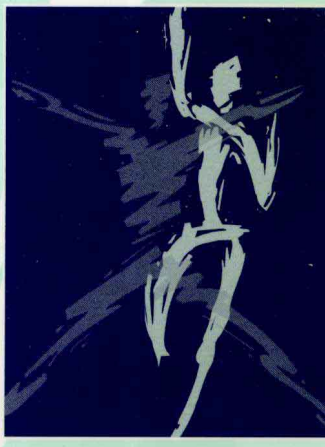
Human Performance Measures Handbook



Valerie J. Gawron

Human Performance Measures Handbook

Valerie J. Gawron
Veridian Engineering

LEA LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
2000 Mahwah, New Jersey London

The material on Measures for Human Performance is based on Appendix B of ANSI/AIAA G—035-1992, *Guide for Human Performance Measurements*, American Institute of Aeronautics and Astronautics, Reston, VA, copyright, 1993.

The final camera copy for this work was prepared by the author, and therefore the publisher takes no responsibility for consistency or correctness of typographical style. However, this arrangement helps to make publication of this kind of scholarship possible.

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, New Jersey 07430-2262

Cover design by Kathryn Houghtaling Lacey

Library of Congress Cataloging-in-Publication Data

Gawron, Valerie, J.
Human performance measures handbook / Valerie J. Gawron.
p. cm.
Includes bibliographical references and indexes.
ISBN 0-8058-3701-9 (pbk. : alk. paper)
1. Human engineering—Handbooks, manuals, etc. 2. Human–machine systems—Handbooks, manuals, etc. I. Title.

T59.7.G38 2000
620.8'2—dc21 00-028774

Books published by Lawrence Erlbaum Associates are printed on acid-free paper, and their bindings are chosen for strength and durability

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Dedication

To my Dad, Stanley C. Gawron, 17 March 1921 to 9 February 2000.

Preface

This human performance measures handbook was developed to help researchers and practitioners select measures to be used in the evaluation of human/machine systems. It can also be used to supplement classes at both the undergraduate and graduate courses in ergonomics, experimental psychology, human factors, human performance, measurement, and system test and evaluation. The handbook begins with an overview of the steps involved in developing a test to measure human performance, workload, and/or situational awareness. This is followed by a definition of human performance and a review of human performance measures. Workload and Situational Awareness are similarly treated in subsequent chapters.

Acknowledgments

This book began while I was supporting numerous test and evaluation projects of military and commercial transportation systems. Working with engineers, operators, managers, programmers, and scientists showed a need for both educating them on human performance measurement and providing guidance for selecting the best measures for the test. I thank my team members for their patience and openness. I also thank Dr. Dave Meister who provided great encouragement to me to write this book based on his reading of my "measure of the month" article in the Test and Evaluation Technical Group newsletter. He and Dr. Tom Enderwick also provided a thorough review of the first draft of this book. For these reviews I am truly grateful.

Contents

List of Figures

List of Tables

1 Introduction

Human factors specialists, including industrial engineers, engineering psychologists, human factors engineers, and many others, consummately seek better (more efficient and effective) ways to characterize and measure the human element as part of the system so we can build trains, planes, and automobiles with superior human–system interfaces. Yet the human factors specialist is often frustrated by the lack of readily accessible information on human performance, workload, and Situational Awareness (SA) measures. This book guides the reader through the critical process of selecting the appropriate measures of human performance, workload, and SA and later, provides specific examples of such.

There are two types of evaluations of human performance. The first type is subjective methods. These are characterized by humans providing opinions through interviews and questionnaires or by observing others' behavior. There are several excellent references on these techniques (e.g., Meister, 1986). The second type of evaluation of human performance is the experimental method. Again there are several excellent references (e.g., Keppel, 1991; Kirk, 1995). This experimental method is the focus of this book. Chapter 1 is a short tutorial on the experimental design; Chapter 2 describes measures of human performance; Chapter 3, measures of workload; and Chapter 4, measures of SA.

For the tutorial, the task of selecting between aircraft cockpit displays is used as an example. For readers familiar with the general principles of experimentation, this should be simply an interesting application of academic theory. For readers who may not be so familiar, it should provide a good foundation of why it is so important to select the right measures in preparation of carrying out your experiment.

The need for efficient and effective selection of the appropriate human performance, workload, and SA measures has never been greater. However, little guidance has been provided to support this selection process. This book was written to meet this need. The book begins with an example in which an experimenter must select measures of performance and workload to evaluate a cockpit display. Next, human performance is defined and measures presented. Each measure is described, along with its strengths and limitations, data requirements, threshold values, and sources of further information. After all the performance measures are described, a procedure for selecting among them is presented. In the last section, workload is defined and workload measures described in the same format as performance measures. To make this desk reference easier to use, extensive author and subjective indices are provided.

1.1 The Example

An experiment is a comparison of two or more ways of doing things. The "things" being done are called *independent variables*. The "ways" of doing things are called *experimental conditions*. The measures used for comparison are *dependent variables*. Designing an experiment requires: defining the independent variables, developing the experimental conditions, and selecting the dependent variables. Ways of meeting these requirements are described in the following steps.

Step 1: Define the Question

Clearly define the question to be answered by the results of the experiment. Let's work through an example. Suppose a moving map display is being designed and the lead engineer wants to know if the map should be designed as track up, north up, or something else. He comes to you for an answer. You have an opinion but no hard evidence. You decide to run an experiment. Start by working with the lead engineer to define the question. First, what are the ways of displaying navigation information, that is, what are the experimental conditions to be compared? The lead engineer responds, "Track up, north up, and maybe something else". If he can not define something else, you can not test it. So now you have two experimental conditions: track up versus north up. These conditions form the two levels of your first independent variable, direction of map movement.

Step 2: Check for Qualifiers

Qualifiers are independent variables that qualify or restrict the generalizability of your results. In our example, an important qualifier is the type of user of the moving map display. Will the user be a pilot (who is used to track up) or a navigator (who has been trained with north-up displays)? If you run the experiment with pilots, the most you can say from your results is that one type of display is best *for pilots*. There is your qualifier. If your lead engineer is designing moving map displays for both pilots and navigators, you have only given him half an answer; or worse, if you did not think about the qualifier of type of user, you may have given him an incorrect answer. So check for qualifiers and use the ones that will have an effect on decision making, as independent variables.

In our example, the type of user will have an effect on decision making, so it should be the second independent variable in the experiment. Also in our example, the size of the display will not have an effect on decision making since the lead engineer only has room for an 8-inch display in the instrument panel. Therefore, size of the display should not be included as an independent variable.

Step 3: Specify Conditions

Specify the exact conditions to be compared. In our example, the lead engineer is interested in track up versus north up. So the movement of the map will vary between the two conditions, but everything else about the displays (e.g., scale factor, display resolution, color quality, size of the display, and so forth) should be exactly the same. This way, if the subjects' performance using the two types of displays is different, that difference can be attributed only to the type of display and not to some other difference between the displays.

Step 4: Match Subjects

Match the subjects to the end users. If you want to generalize the results of your experiment to what will happen in the real world, try to match the subjects to the users of the system in the real world. This is extremely important since subjects' past experiences may greatly affect their performance in an experiment. In our example, we added a second independent variable to our experiment specifically because of subjects' previous experiences (that is, pilots are used to track up, navigators are trained with north up). If the end users of the display are pilots, we should use pilots as our subjects. If the end users are navigators, we should use navigators as our subjects. Other subject variables may also be important; in our example, age and training are both very important. Therefore, you should identify what training the user of the map display must have and provide that same training to the subjects before the start of data collection.

Age is important because pilots in their 40s may have problems focusing on near objects such as map displays. Previous training is also important: F-16 pilots have already used moving map displays while C-130 pilots have not. If the end users are pilots in their 20s with F-16 experience, and your subjects are pilots in their forties with C-130 experience, you may be giving the lead engineer the wrong answer to his question of which type of display is better.

Step 5: Select Performance Measures

Your results are influenced to a large degree by the performance measures you select. Performance measures should be relevant, reliable, valid, quantitative, and comprehensive. Let's use these criteria to select performance measures for our example problem.

Criteria 1: Relevant. Relevance to the question being asked is the prime criteria to be used when selecting performance measures. In our example, the lead engineer's question is "What type of display format is better?" Better can refer to staying on course better (accuracy) but it can also refer to getting to the waypoints on time better (time). Subjects' ratings of which display format they prefer does not answer the question of which display

is better from a performance standpoint because preference ratings can be affected by factors other than performance.

Criteria 2: Reliable. Reliability refers to the repeatability of the measurements. For recording equipment, reliability is dependent on careful calibration of equipment to ensure that measurements are repeatable and accurate; (i.e., an actual course deviation of 50.31 feet should always be recorded as 50.31 feet). For rating scales, reliability is dependent on the clarity of the wording. Rating scales with ambiguous wording will not give reliable measures of performance. For example, if the question on the rating scale is "Was your performance okay?" the subject may respond "No" after his first simulated flight but "Yes" after his second, simply because he is more comfortable with the task. If you now let him repeat his first flight, he may respond, "Yes." In this case, you are getting a different answer to the same question in the same condition. Subjects will give more reliable responses to less ambiguous questions such as "Did you deviate more than 100 feet from course in this trial?" Even so, you may still get a first "No" and a second "Yes" to the more precise question, indicating that some learning had improved his performance the second time.

Subjects also need to be calibrated. For example, if you are asking which of eight flight control systems is best and your metric is an absolute rating (e.g., Cooper-Harper Handling Qualities Rating), your subject needs to be calibrated with both a "good" aircraft and a "bad" aircraft at the beginning of the experiment. He may also need to be recalibrated during the course of the experiment. The symptoms that suggest the need to recalibrate your subject are the same as those that indicate that you should recalibrate your measuring equipment: (a) all the ratings are falling in a narrower band than you expect, (b) all the ratings are higher or lower than you expect, and (c) the ratings are generally increasing (or decreasing) across the experiment independent of experimental condition. In these cases, give the subject a flight control system that he has already rated. If this second rating is substantially different from the one he previously gave you for the same flight control system, you need to recalibrate your subjects with an aircraft that pulls their ratings away from the average: bad aircraft if all the ratings are near the top, good aircraft if all the ratings are near the bottom.

Criteria 3: Valid. Validity refers to measuring what you really think you are measuring. Validity is closely tied to reliability. If a measure is not reliable, it can never be valid. The converse is not necessarily true. For example, if you ask a subject to rate his workload from 1 to 10 but do not define for him what you mean by workload, he may rate the perceived difficulty of the task rather than the amount of effort he expended in performing the task.

Criteria 4: Quantitative. Quantitative measures are easier to analyze than qualitative measures. They also provide an estimate of the size of the difference between experimental conditions. This is often very useful in performing trade-off analyses of performance versus cost of system designs. This criterion does not preclude the use of qualitative measures, however, because qualitative measures often improve the understanding of experiment results. For qualitative measures, an additional issue must be considered - the type of rating scale. Nominal scales assign an adjective to system being evaluated, (e.g., easy to use). "A nominal scale is categorical in nature, simply identifying differences among things on some characteristic. There is no notion of order, magnitude or size" (Morrow, Jackson, Disch, and Mood, 1995, p. 28). Ordinal scales

rank systems being evaluated on a single or a set of dimensions (e.g., the north-up is easier than the track-up display). "Things are ranked in order, but the difference between ranked positions are not comparable" (Morrow, Jackson, Disch, and Mood, 1995, p. 28). Interval scales have equal distances between the values being used to rate the system under evaluation. For example, a bipolar rating scale is used in which the two poles are *extremely easy to use* and *extremely difficult to use*. In between these extremes are the words *moderately easy, equally easy,* and *moderately difficult*. The judgment is that there is an equal distance between any two points on the scale. The perceived difficulty difference between *extremely* and *moderately* is the same as between *moderately* and *no difference*. However, "the zero point is arbitrarily chosen" (Morrow, Jackson, Disch, and Mood, 1995, p. 28). The fourth type of scale is a ratio scale which possesses a true zero (Morrow, Jackson, Disch, and Mood, 1995, p. 29). More detailed descriptions of scales are presented in Baird and Noma (1978), Torgerson (1958), and Young (1984).

Criteria 5: Comprehensive. Comprehensive means the ability to measure all aspects of performance. Recording multiple measures of performance during an experiment is cheaper than setting up a second experiment to measure something that you missed in the first experiment. So measure all aspects of performance that may be influenced by the independent variables. In our example, subjects can trade off accuracy for time (e.g., cut a leg to reach a waypoint on time) and vice versa (e.g., go slower to stay on course better), so we should record both accuracy and time measures.

Step 6: Use Enough Subjects

Use enough subjects to statistically determine if there is a difference in the values of the dependent variables between the experimental conditions. In our example, is the performance of subjects using the track-up display versus the north-up display statistically different? Calculating the number of subjects you need is very simple. First, predict how well subjects will perform in each condition. You can do this using your own judgment, previous data from similar experiments, or from pretest data using your experimental setup. In our example, how much error will there be in waypoint arrival times using the track-up display and the north-up display? From previous studies, you may think that the average error for pilots using the track-up display will be 1.5 seconds and using the north-up display, 2 seconds. Similarly, the navigators will have about 2 seconds error using the track-up display and 1.5 seconds error with the north-up display. For both sets of subjects and both types of displays, you think the standard deviation will be about 0.5 second.

Now we can calculate the effect size, that is, the difference between performances in each condition:

$$\text{effect size} = \frac{|\text{performance in track up} - \text{performance in north up}|}{\text{standard deviation}}$$

$$\text{effect size for pilots} = \frac{|1.5 - 2|}{0.5} = 1$$

$$\text{effect size for navigators} = \frac{|2 - 1.5|}{0.5} = 1$$