

HUMAN PERFORMANCE MEASURES HANDBOOK

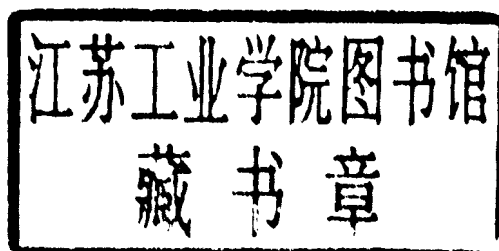
VALERIE J. GAWRON



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.

Copyright © 2000 by Lawrence Erlbaum Associates, Inc.

All rights reserved. No part of this book may be reproduced in any form, by photostat, microfilm, retrieval system, or any other means, without the prior written permission of the publisher.

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	xi
List of Tables	xii
Preface.....	xv
Acknowledgments.....	xv
1 Introduction.....	1
1.1 The Example	2
Step 1: Define the Question	2
Step 2: Check for Qualifiers.....	2
Step 3: Specify Conditions.....	3
Step 4: Match Subjects.....	3
Step 5: Select Performance Measures	3
Step 6: Use Enough Subjects	5
Step 7: Select Data-Collection Equipment.....	6
Step 8: Match Trials.....	7
Step 9: Select Data-Recording Equipment	8
Step 10: Decide Subject Participation.....	8
Step 11: Order the Trials.....	9
Step 12: Check for Range Effects	9
1.2 Summary	10
1.3 References.....	11
2 Human Performance	13
2.1 Accuracy	14
2.1.1 Absolute Error.....	14
2.1.2 Average Range Score.....	14
2.1.3 Deviations	14
2.1.4 Error Rate.....	14
2.1.5 False Alarm Rate.....	15
2.1.6 Number Correct	15
2.1.7 Number of Errors	15
2.1.8 Percent Correct.....	16
2.1.9 Percent Errors.....	17
2.1.10 Probability of Correct Detections	17
2.1.11 Ratio of Number Correct/Number Errors	18
2.1.12 Root Mean Square Error	18
2.2 AGARD's Standardized Tests for Research with Environmental Stressors (STRES) Battery	20
2.3 Aircraft Parameters	21
2.3.1 Takeoff and Climb	21
2.3.2 Cruise	21
2.3.3 Approach and Landing.....	22
2.3.4 Hover.....	23
2.4 Armed Forces Qualification Test.....	25
2.5 Boyett and Conn's White-Collar Performance Measures.....	25
2.6 Charlton's Measures of Human Performance in Space Control Systems	27

2.7	Control Input Activity.....	27
2.8	Correctness Score	28
2.9	Critical Incident Technique	28
2.10	Deutsch and Malmberg Measurement Instrument Matrix.....	29
2.11	Dichotic Listening.....	29
2.12	Driving Parameters	30
2.12.1	Average Brake RT	30
2.12.2	Brake Pedal Errors	31
2.12.3	Control Light Response Time.....	31
2.12.4	Number of Brake Responses.....	31
2.12.5	Perception-Response Time	31
2.12.6	Speed.....	31
2.12.7	Steering Wheel Reversals	31
2.12.8	Time	32
2.12.9	Tracking Error.....	32
2.13	Eastman Kodak Company Measures for Handling Tasks	34
2.14	Glance	34
2.15	Haworth-Newman Avionics Display Readability Scale.....	35
2.16	Landing Performance Score.....	35
2.17	Lookpoint.....	37
2.18	Marking Speed and Errors	37
2.19	Mental Arithmetic.....	38
2.20	Movement Time.....	38
2.21	Nieva, Fleishman, and Rieck's Team Dimensions	39
2.22	Performance Evaluation Tests for Environmental Research (PETER)	39
2.23	Pilot Performance Index	40
2.24	Reaction Time.....	41
2.24.1	Auditory Stimuli	41
2.24.2	Tactile Stimuli.....	42
2.24.3	Visual Stimuli	42
2.24.4	Related Measures	44
2.25	Reading Speed	47
2.26	Search Time	49
2.27	Simulated Work and Fatigue Test Battery.....	50
2.28	Task Load	50
2.29	Time to Complete	51
2.30	Time-to-Line-Crossing (TLC)	52
2.31	Unified Tri-services Cognitive Performance Assessment Battery (UTCPAB)...	53
3	Human Workload.....	54
3.1	Performance Measures of Workload	55
3.1.1	Aircrew Workload Assessment System.....	55
3.1.2	Control Movements/Unit Time.....	55
3.1.3	Glance Duration and Frequency	56
3.1.4	Load Stress.....	57
3.1.5	Observed Workload Area.....	57
3.1.6	Rate of Gain of Information.....	58

3.1.7 Relative Condition Efficiency.....	58
3.1.8 Speed Stress	58
3.1.9 Secondary Tasks	59
3.1.9.1 Card Sorting Secondary Task	60
3.1.9.2 Choice RT Secondary Task	61
3.1.9.3 Classification Secondary Task.....	64
3.1.9.4 Cross-Adaptive Loading Secondary Task	65
3.1.9.5 Detection Secondary Task	65
3.1.9.6 Distraction Secondary Task	66
3.1.9.7 Driving Secondary Task	67
3.1.9.8 Identification/Shadowing Secondary Task	68
3.1.9.9 Lexical Decision Secondary Task.....	69
3.1.9.10 Memory-Scanning Secondary Task.....	70
3.1.9.11 Mental Mathematics Secondary Task.....	73
3.1.9.12 Michon Interval Production Secondary Task	76
3.1.9.13 Monitoring Secondary Task.....	78
3.1.9.14 Multiple Task Performance Battery of Secondary Tasks	82
3.1.9.15 Occlusion Secondary Task.....	83
3.1.9.16 Problem-Solving Secondary Task.....	84
3.1.9.17 Production/Handwriting Secondary Task.....	85
3.1.9.18 Psychomotor Secondary Task.....	85
3.1.9.19 Randomization Secondary Task	86
3.1.9.20 Reading Secondary Task.....	87
3.1.9.21 Simple Reaction-Time Secondary Task.....	87
3.1.9.22 Simulated Flight Secondary Task	89
3.1.9.23 Spatial-Transformation Secondary Task.....	89
3.1.9.24 Speed-Maintenance Secondary Task	90
3.1.9.25 Sternberg Memory Secondary Task.....	90
3.1.9.26 Three-Phase Code Transformation Secondary Task.....	94
3.1.9.27 Time-Estimation Secondary Task.....	95
3.1.9.28 Tracking Secondary Task	97
3.1.9.29 Workload Scale Secondary Task	100
3.1.10 Task Difficulty Index	101
3.1.11 Time Margin	101
3.2 Subjective Measures of Workload	102
3.2.1 Analytical Hierarchy Process.....	104
3.2.2 Arbeitswissenschaftliches Erhebungsverfahren zur Tätigkeitsanalyse	106
3.2.3 Bedford Workload Scale.....	106
3.2.4 Computerized Rapid Analysis of Workload	108
3.2.5 Continuous Subjective Assessment of Workload	109
3.2.6 Cooper-Harper Rating Scale	109
3.2.7 Crew Status Survey	111
3.2.8 Dynamic Workload Scale	114
3.2.9 Equal-Appearing Intervals	114
3.2.10 Finegold Workload Rating Scale	115
3.2.11 Flight Workload Questionnaire.....	116

3.2.12	Hart and Bortolussi Rating Scale.....	116
3.2.13	Hart and Hauser Rating Scale.....	117
3.2.14	Honeywell Cooper-Harper Rating Scale	118
3.2.15	Magnitude Estimation.....	119
3.2.16	McCracken-Aldrich Technique.....	120
3.2.17	McDonnell Rating Scale.....	120
3.2.18	Mission Operability Assessment Technique.....	120
3.2.19	Modified Cooper-Harper Rating Scale	122
3.2.20	Multi-Descriptor Scale.....	126
3.2.21	Multidimensional Rating Scale.....	126
3.2.22	NASA Bipolar Rating Scale	127
3.2.23	NASA Task Load Index.....	130
3.2.24	Overall Workload Scale.....	135
3.2.25	Pilot Objective/Subjective Workload Assessment Technique.....	136
3.2.26	Pilot Subjective Evaluation.....	137
3.2.27	Profile of Mood States	137
3.2.28	Sequential Judgment Scale	140
3.2.29	Subjective Workload Assessment Technique.....	141
3.2.30	Subjective Workload Dominance	149
3.2.31	Task Analysis Workload.....	149
3.2.32	Utilization	150
3.2.33	Workload/Compensation/Interference/Technical Effectiveness.....	151
3.2.34	Zachary/Zaklad Cognitive Analysis.....	152
3.3	Simulation of Workload.....	153
4	Measures of Situational Awareness.....	155
4.1	Performance Measures of SA	157
4.1.1	Situational Awareness Global Assessment Technique	157
4.1.2	Situational Awareness Linked Instances Adapted to Novel Tasks.....	159
4.1.3	Temporal Awareness.....	160
4.2	Subjective Measures of SA.....	160
4.2.1	China Lake Situational Awareness	160
4.2.2	Crew Situational Awareness.....	161
4.2.3	Human Interface Rating and Evaluation System	162
4.2.4	Situational Awareness Rating Technique	162
4.2.5	Situational Awareness Subjective Workload Dominance	165
4.2.6	Situational Awareness Supervisory Rating Form	165
4.3	Simulation.....	165
	Glossary of Terms.....	167
	Author Index	169
	Subject Index	183

List of Figures

FIG. 1. Number of subjects needed as a function of effect size.	6
FIG. 2. Haworth-Newman Display Readability Rating Scale (from Haworth, 1993 cited in Chiappetti, 1994)	36
FIG. 3. Sternberg Memory Task Data	91
FIG. 4. Example AHP Rating Scale.....	104
FIG. 5. Bedford Workload Scale	107
FIG. 6. Cooper-Harper Rating Scale.....	110
FIG. 7. Crew Status Survey	112
FIG. 8. Dynamic Workload Scale.....	114
FIG. 9. Finegold Workload Rating Scale.....	115
FIG. 10. Hart and Hauser Rating Scale.....	117
FIG. 11. Honeywell Cooper-Harper Rating Scale	118
FIG. 12. McDonnell Rating Scale.....	121
FIG. 13. Modified Cooper-Harper Rating Scale.....	123
FIG. 14. NASA Bipolar Rating Scale.....	128
FIG. 15. NASA TLX Rating Sheet.....	130
FIG. 16. Pilot Subjective Evaluation Scale.....	138
FIG. 17. 15-point Form of the Sequential Judgment Scale (Pfender, Pitrella, and Wiegand, 1994, p. 31).	141
FIG. 18. WCI/TE Scale Matrix.....	151
FIG. 19. Decision making under uncertainty and time pressure (Dorfel and Distelmaier, 1997, p. 2)	156
FIG. 20. Guide to selecting a SA measure.....	156
FIG. 21. SART Scale	162

List of Tables

TABLE 1. Component Abilities of Commercial Airline Pilot Performance Determined by Frequency of Errors Extracted from Accident Reports, Critical Incidents, and Flight Checks	21
TABLE 2. White-Collar Measures in Various Functions.....	26
TABLE 3. Pilot Performance Index Variable List.....	40
TABLE 4. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Choice RT Task	62
TABLE 5. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Detection Task.....	66
TABLE 6. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Identification Task.....	68
TABLE 7. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Memory Task.....	71
TABLE 8. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Task	75
TABLE 9. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Michon Interval Production Task	77
TABLE 10. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Monitoring Task	80
TABLE 11. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Occlusion Task	83
TABLE 12. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Problem-Solving Task	84
TABLE 13. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Psychomotor Task	85
TABLE 14. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Randomization Task	86
TABLE 15. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Simple RT Task.....	88
TABLE 16. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Task	92
TABLE 17. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Task	96
TABLE 18. References Listed by the Effect on Performance of Primary Tasks Paired with a Secondary Tracking Task	98
TABLE 19. Comparison of Subjective Measures of Workload	103
TABLE 20. Definitions of AHP Scale Descriptors	105
TABLE 21. Mission Operability Assessment Technique Pilot Workload and Subsystem Technical Effectiveness Rating Scales	122
TABLE 22. Multidimensional Rating Scale.....	126
TABLE 23. NASA Bipolar Rating-Scale Descriptions.....	128

TABLE 24. NASA TLX Rating-Scale Descriptions	131
TABLE 25. SWAT Scales	142
TABLE 26. Generic Behavioral Indicators of Team SA (Muniz, Stout, Bowers, and Salas, 1998)	159
TABLE 27. China Lake SA Rating Scale.....	161
TABLE 28. Definitions of SART Rating Scales	163
TABLE 29. Situational Awareness Supervisory Rating Form	166

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:

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

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

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