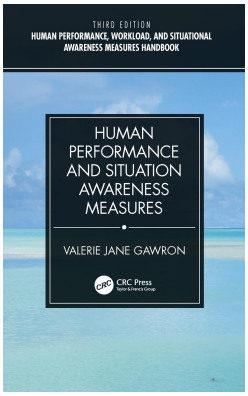

Human Performance and Situation Awareness Measures

Gawron Valerie Jane

Introduction

Publication details
https://test.routledgehandbooks.com/doi/10.1201/9780429001024-1
Gawron Valerie Jane
Published online on: 09 Jan 2019

1

Introduction

Human factors specialists, including ergonomists, industrial engineers, engineering psychologists, human factors engineers, and many others, continually 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; process control stations, and other systems 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. To fill that void, this book was written to guide the reader through the critical process of selecting the appropriate measures of human performance, workload, and SA for objective evaluations.

There are two types of evaluations of human performance. The first type is subjective measures. 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. 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 when preparing to conduct an experiment.

Chapter 2 describes measures of human performance and Chapter 3 describes measures of SA. Each measure is described, along with its strengths and limitations, data requirements, threshold values, and sources of further information. 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.

1.1.1 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 cannot define something else, you cannot 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.

1.1.2 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.

1.1.3 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 participants' 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.

1.1.4 Step 4: Match Participants

Match the participants 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 participants to the users of the system in the real world. This is extremely important since participants' 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 participants' 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 participants. If the end users are navigators, we should use navigators as our participants. Other participant 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 participants before the start of data collection.

Age is important because pilots in their forties 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 twenties with F-16 experience and your participants 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.

1.1.5 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). Participants' 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 participant 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. Participants 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.

Participants 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 participant 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 participant 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 participant 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 participants 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 participant 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 the 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 et al., 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 et al., 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 et al., 1995, p. 28). The final type of scale is a ratio scale which possesses a true zero (Morrow et al., 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, participants 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. For an example of using these and several additional criteria in air combat maneuvering, see Lane (1986).

1.1.6 Step 6: Use Enough Participants

Use enough participants 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 participants using the track-up display versus the north-up display statistically different? Calculating the number of participants you need is very simple. First, predict how well participants 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 participants 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:

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

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

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

In Figure 1.1 we can now read the number of participants needed to discriminate the two conditions. For an effect size of 1, the number of participants needed is 18. Therefore, we need 18 pilots and 18 navigators in our experiment. Note that although the function presented in Figure 1.1 is not etched in stone, it is based on over 100 years of experimentation and statistics.

Note that you should estimate your effect size in the same units as you will use in the experiment. Also note that because effect size is calculated as a ratio, you will get the same effect size (and hence the same number of participants) for equivalent measures. Finally, if you have no idea of the effect size, try the experiment yourself and use your own data to estimate the effect size.

1.1.7 Step 7: Select Data-Collection Equipment

Now that you know the size of the effect of the difference between conditions, check that the data-collection equipment you have selected can reliably measure performance at least one order of magnitude smaller than the smallest discriminating decimal place in the size of the expected difference between conditions. In our example, the expected size in one condition was 1.5 seconds. The smallest discriminating decimal place (1.5 vs. 2.0) is tenths of a second. One order of magnitude smaller is hundredths. Therefore, the recording equipment should be accurate to 1/100th of a second.

1.1.8 Step 8: Match Trials

Match the experimental trials to the end usage. As in Step 4, if you want to generalize the results of your experiment to what will happen in the real world, you must match the experimental trials to the real world.

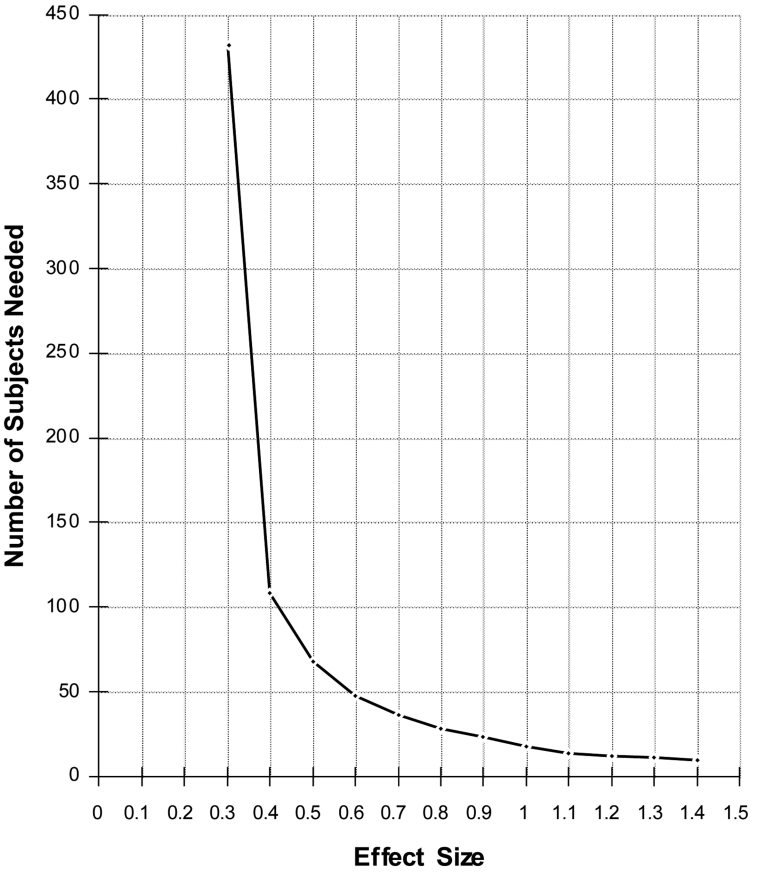

FIGURE 1.1
Number of participants needed as a function of effect size.

(Note, a single trial is defined as continuous data collection under the same experimental conditions. For example, three successive instrument approaches with the same flight-control configuration constitute one trial.) The following are important characteristics to match.

Characteristic 1: Length of the Trial. Over the length of a trial, performance improves due to warm-up effects and learning and then degrades as fatigue sets in. If you measure performance in the experiment for 10 minutes but in the real world, pilots and navigators perform the task for two hours, your results may not reflect the peak warm-up or the peak fatigue. Consequently, you may give the lead engineer the wrong answer. So always try to match the length of each experimental trial to the length of the task in the real world.

Characteristic 2: Level of Difficulty. If you make the experimental task too easy, all the participants will get the same performance score: 0 errors. If all the performance scores are the same, you will not be able to distinguish

between experimental conditions. To avoid this problem, make the task realistically difficult. In general, the more difficult the task in the experiment, the more likely you are to find a statistical difference between experimental conditions. This is because difficulty enhances discriminability between experimental conditions. However, there are two exceptions that should be avoided in any experiment. First, if the experimental task is too difficult, the performance of all the participants will be exactly the same: 100% errors. You will have no way of knowing which experimental condition is better and the experiment was useless. Second, if you increase the difficulty of the task beyond that which can ever be expected in the real world, you may have biased your results. In our example, you may have found that track-up displays are better than north-up displays in mountainous terrain, flying under 100 feet Above Ground Level (AGL) at speeds exceeding 500 knots with wind gusts over 60 knots. But how are they in hilly terrain, flying at 1000 feet AGL at 200 knots with wind gusts between 10 and 20 knots, that is, in the conditions in which they will be used nearly 70% of the time? You cannot answer this question from the results of your experiment – or if you give an answer, it may be incorrect. Therefore, typical conditions should be conditions of the experiment.

Characteristic 3: Environmental Conditions. Just as in Step 4 where you tried to match the participants to the end users, you should also try to match the environmental conditions of the laboratory (even if that laboratory is an operational aircraft or an in-flight simulator) to the environmental conditions of the real world. This is extremely important because environmental conditions can have a greater effect on performance than the independent variables in your experiment. Important environmental conditions that should be matched include lighting, temperature, noise, and task load. Lighting conditions should be matched in luminance level (possible acuity differences), position of the light source (possible glare), and type of light source (incandescent lights have "hot spots" that can create point sources of glare; fluorescent lights provide even, moderate light levels; sunlight can mask some colors and create large glare spots). Temperatures above 80 degrees Fahrenheit decrease the amount of effort participants expend; temperatures below 30 degrees Fahrenheit make fine motor movements (e.g., setting radio frequencies) difficult. Noise can either enhance or degrade performance: enhancements are due to increased attention; degradations are due to distractions. Meaningful noise (e.g., a conversation) is especially distracting. Task load refers to both the number and types of tasks that are being performed at the same time as your experimental task. In general, the greater the number of tasks that are being performed simultaneously and the greater the similarity of the tasks that are being performed simultaneously, the worse the performance on the experimental task. The classic example is monitoring three radio channels simultaneously. If the volume or quality of the communications is not varied (thus making the tasks less similar), this task is extremely difficult.

1.1.9 Step 9: Select Data-Recording Equipment

In general, the data-recording equipment should be able to record data for 1.5 times the length of the experimental trial. This allows for false starts without changing the data tape, disk, or other storage medium. The equipment should be able to have separate channels for each continuous dependent variable (e.g., altitude, airspeed) and as many channels as necessary to record the discrete variables (e.g., reaction time (RT) to a simulated fire) without any possibility of recording the discrete variables simultaneously on the same channel (thus losing valuable data).

1.1.10 Step 10: Decide Participant Participation

Decide if each participant should participate in all levels of the condition in the experiment. There are many advantages of having a single participant participate in more than one experimental condition: (a) reduced recruitment costs, (b) decreased total training time, and (c) better matching of participants across experimental conditions. But there are some conditions that preclude using the same participant in more than one experimental condition. The first is previous training. In our example, pilots and navigators have had very different training. The differences in their training may affect their performance; therefore, they cannot participate in both roles: pilot and navigator. Second, some experimental conditions can make the participants' performance worse than even untrained participants in another experimental condition. This effect is called negative transfer. Negative transfer is especially strong when two experimental conditions require a participant to give a different response to the same stimulus. For example, the response to a fire alarm in Experimental Condition 1 is pull the T handles, then feather the engine. In Experimental Condition 2, the response is feather the engine and then pull the T handle. Participants who have not participated in any experimental condition are going to have faster RTs and fewer errors than participants who have already participated in either Experimental Condition 1 or 2. Whenever there is negative transfer (easy to find by comparing performance of new participants to participants who have already participated in another condition), use separate participants.

Learning is another important condition affecting the decision to use the same participants or not. Participants who participate in more than one experimental condition are constantly learning about the task that they are performing. If you plot the participants' performance (where high scores mean good performance) on the ordinate and the number of trials he/she has completed along the abscissa, you will find a resulting J curve where a lot of improvement in performance occurs in the first few trials and very little improvement occurs in the final trials. The point at which there is very little improvement is called *asymptotic learning*. Unless participants are all trained to asymptote before the first trial, their performance will improve

over the entire experiment regardless of the differences in the experimental conditions. Therefore, the "improvement" you see in later experimental conditions may have nothing to do with what the experimental condition is but rather with how long the participant has been performing the task in the entire experiment.

A similar effect occurs in simple, repetitive, mental tasks and all physically demanding tasks. This effect is called *warm-up*. If the participants' performance improves over trials regardless of the experimental conditions, you may have a warm-up effect. This effect can be eliminated by having participants perform preliminary trials until their performance on the task matches their asymptotic learning.

The final condition is fatigue. If the same participant is performing more than one trial, fatigue effects may begin to mask the differences in the experimental conditions. You can check for fatigue effects in four ways: by performing a number of trials yourself (how are you feeling?); by observing your participants (are they showing signs of fatigue?); by comparing performance in the same trial number but different conditions across participants (is everyone doing poorly after three trials?); and by asking the participants how they are feeling.

1.1.11 Step 11: Order the Trials

In Step 10, we described order or carry over effects. Even if these do not occur to a great degree or if they do not seem to occur at all, it is still important to order your data-collection trials so as to minimize order and carry over effects. Another important carry over effect is the experimenter's experience – during your first trial, experimental procedures may not yet be smoothed out. By the 10th trial, everything should be running efficiently and you may even be anticipating participants' questions before they ask them. The best way to minimize order and carry over effects is to use a Latin-square design. This design ensures that every experimental condition precedes and succeeds every other experimental condition an equal number of times.

Once the Latin square is generated, check the order for any safety constraints (e.g., landing a Level 3 aircraft in maximum turbulence or severe crosswinds). Adjust this order as necessary to maintain safety. The resulting numbers indicate the order in which you should collect your data. For example, Participant 1 gets north up then track up. Participant 2 gets the opposite. Once you have completed data collection for the pilots, you can collect data on the navigators. It does not matter what order you collect the pilots' and navigators' data because the pilots' data will never be compared to the navigators' data, that is, you are not looking for an interaction between the two independent variables. If the second independent variable in the experiment had been size (e.g., the lead engineer gives you the option for an 8- or 12-inch display), the interaction would have been of interest. For example, are 12-inch, track-up displays better than 8-inch, north-up

displays? If we had been interested in this interaction, a Latin square for four conditions: Condition 1, 8-inch, north up; Condition 2, 8-inch, track up; Condition 3, 12-inch, north up; and Condition 4, 12-inch, track up would have been used.

1.1.12 Step 12: Check for Range Effects

Range effects occur when your results differ based on the range of experimental conditions that you use. For example, in Experiment 1 you compare track-up and north-up displays, and find that for pilots track-up displays are better. In Experiment 2, you compare track-up, north-up, and horizontal situation indicator (HSI) displays. This time you find no difference between track-up and north-up displays but both are better than a conventional HSI. This is an example of a range effect: when you compare across one range of conditions, you get one answer; when you compare across a second range of conditions, you get another answer. Range effects are especially prevalent when you vary environmental conditions such as noise level and temperature. Range effects cannot be eliminated. This makes selecting a range of conditions for your experiment especially important.

To select a range of conditions, first return to your original question. If the lead engineer is asking which of two displays to use, Experiment 1 is the right experiment. If he is asking whether track-up or north-up displays are better than an HSI, Experiment 2 is correct. Second, you have to consider how many experimental conditions your participants are experiencing. If it is more than seven, your participant is going to have a hard time remembering what each condition was but his or her performance will still show the effect. To check for a "number of trials" effect, plot the average performance in each trial versus the number of the trials the participant has completed. If you find a general decrease in performance, it is time to either reduce the number of experimental conditions that the participant experiences or provide long rest periods.

1.2 Summary

The quality and validity of the data are improved by incorporating the following steps in the experimental design:

Step 1: Clearly define the question to be answered.

Step 2: Check for qualifiers.

Step 3: Specify the exact conditions to be compared.

Step 4: Match the participants to the end users.

Step 5: Select performance measures.
Step 6: Use enough participants.
Step 7: Select data-collection equipment.
Step 8: Match the experimental trials to the end usage.
Step 9: Select data-recording equipment.
Step 10: Decide if each participant should participate in all levels.
Step 11: Order the trials.
Step 12: Check for range effects.

Step 5 is the focus for the remainder of this book.

Sources

Baird, J.C., and Noma, E. *Fundamentals of Scaling and Psychophysics*. New York: Wiley, 1978.

Keppel, G. *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs, NJ: Prentice Hall, 1991.

Kirk, R.R. *Experimental Design: Procedures for the Behavioral Sciences*. Pacific Grove, California: Brooks/Cole Publishing Company, 1995.

Lane, N. *Issues in Performance Measurement for Military Aviation with Applications to Air Combat Maneuvering (NTSC TR-86-008)*. Orlando: Naval Training Systems Center, April 1986.

Meister, D. *Human Factors Testing and Evaluation*. New York: Elsevier, 1986.

Morrow, J.R., Jackson, A.W., Disch, J.G., and Mood, D.P. *Measurement and Evaluation in Human Performance.* Champaign, IL: Human Kinematics, 1995.

Torgerson, W.S. *Theory and Methods of Scaling.* New York: Wiley, 1958.

Young, F.W. Scaling. *Annual Review of Psychology* 35: 55–81, 1984.