nngroup.com

How to Conduct a Heuristic Evaluation

Kate Moran, Kelley Gordon

10–13 minutes

Summary: Step-by-step instructions to systematically review your product to find potential usability and experience problems. Download a free heuristic evaluation template.

A **heuristic evaluation** is a method for identifying design problems in a user interface. Evaluators judge the design against a set of guidelines (called heuristics) that make systems easy to use.

(For more information about how and why this tool was developed, read Jakob Nielsen's 1994 article, "The Theory Behind Heuristic Evaluations".)

- References

Choosing a Set of Heuristics

A heuristic evaluation can be conducted with any set of heuristics. To assess usability, we recommend Jakob Nielsen's 10 usability heuristics — a set of high-level guidelines based on an understanding of human behavior, psychology, and information processing. For specialized domains or types of usability assessments, you may consider using other domain-specific ones in addition.

When to Conduct a Heuristic Evaluation

Heuristic evaluations are useful for identifying glaring problems in an interface. That interface can be just about anything that users will interact with — including prototypes, physical products, games, virtual reality, or voice interfaces. The method can be particularly helpful early in the design process.

Heuristic evaluations are **useful for stretching a limited UX research budget,** because they help you find likely issues without having to test with participants.

However, **heuristic evaluations cannot replace user research**. User-experience design is highly contextual. To design good experiences, you'll still need to test with actual users. But heuristic evaluations can complement your team's research work; for example, conducting a heuristic evaluation in preparation for an upcoming usability test might help you identify the elements of the design that you should target during testing.

Conducting heuristic evaluations is also a good way to **develop strong UX instincts.** If you're new to UX, consider using heuristic evaluations as a way to train yourself to catch common usability issues. Practice conducting these evaluations on many different types of products — whether you actually work on them or not.

Step 1: Prepare for a Heuristic Evaluation

Choose and Train Your Team

Heuristic evaluations work best when performed by a group of people, not just by one evaluator. This is because each individual (no matter how experienced or expert) is likely to miss some of the potential usability issues. Ideally, **three to five people should independently evaluate** the same interface.

Teams conducting their first heuristic evaluation will need a bit of **training and preparation** before they begin. Start by asking each person to read and understand the heuristics.

Next, consider doing a **practice round** with a simple design as a group. You might conduct a collaborative evaluation of a weather app, for example. The point of this practice round is to ensure that everyone on the team understands what they're expected to do during an evaluation (more details on this in step 2.)

Decide How to Document Evaluations

Your evaluators will need a place to collect their observations. You might use:

- **Our heuristic-evaluation workbook**: Each team member can fill out a printed or digital version of this interactive PDF. Download our free workbook to use it in your evaluation.
- **A spreadsheet:** Evaluators can capture one observation per line, along with its corresponding heuristic.

- **A digital whiteboard:** In a tool like Miro or Mural, create separate workspaces for each evaluator. Include screenshots of the interface and have evaluators place sticky notes directly on the elements they're analyzing.

  If you choose to use a shared document or space (like Google Sheets or a digital-whiteboard tool), **your team members should not see each other's evaluations until their own evaluation is complete**. The point of having multiple evaluators is to capture independent observations, so you don't want team members to influence each other.

Set the Scope

The **narrower the scope,** the easier and more detailed the evaluation will be. For your team's first heuristic evaluation — or if you have a large, complex product — consider keeping your scope narrow to make things manageable.

Narrow your scope by looking at:

- One task at a time
- One section of the site or app
- One user group, if you have many with diverse needs
- One device type (desktop, tablet, mobile)

Step 2: Evaluate Independently

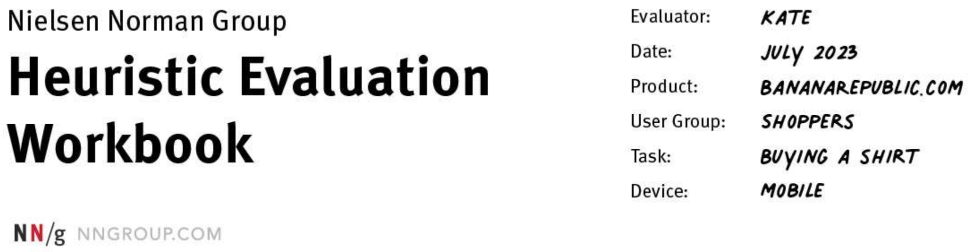
Next, each team member should evaluate the interface on their own.

It's important to timebox this activity to make it manageable. We recommend reserving about **1–2 hours.**

Become Familiar with the Product

Let’s consider a simple ecommerce example to explain how the evaluation might work.

Product: Banana Republic site
User Group: Shoppers
Task: Buying a shirt
Device: Mobile



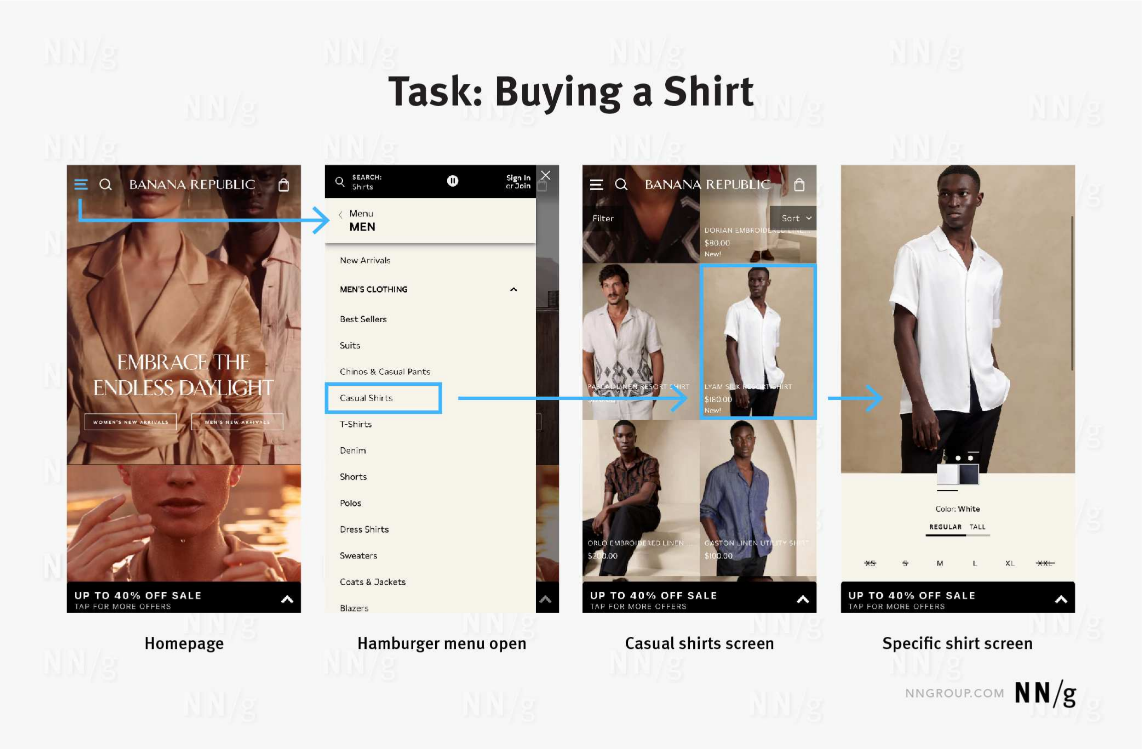
Nielsen Norman Group

Heuristic Evaluation Workbook

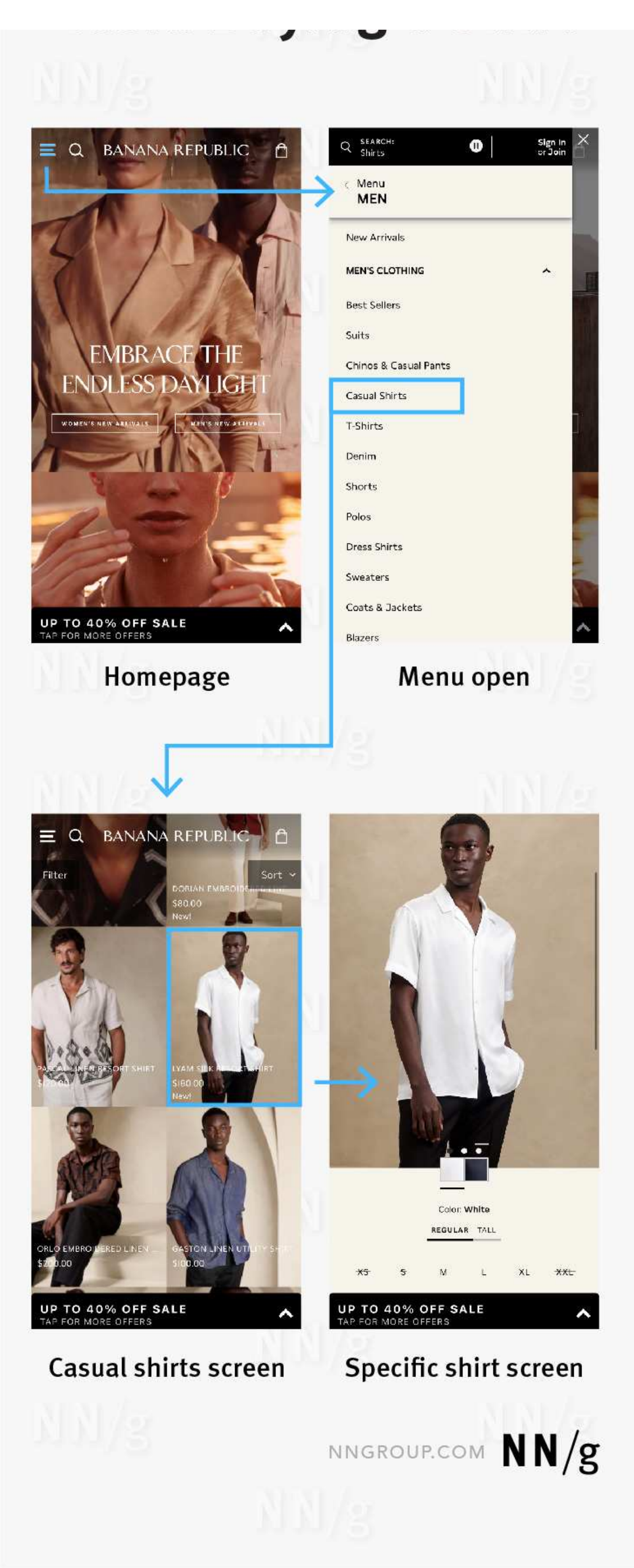
NN/g NNGROUP.COM

Evaluator: KATE
Date: JULY 2023
Product: BANANAREPUBLIC.COM
User Group: SHOPPERS
Task: BUYING A SHIRT
Device: MOBILE

Fill in the details of the product you'll be reviewing at the topc of your workbook. Include the specific user group, task, or device, if you've narrowed your scope.



Bananarepublic.com: A possible flow leading to a product page



Bananarepublic.com: A possible flow leading to a product page

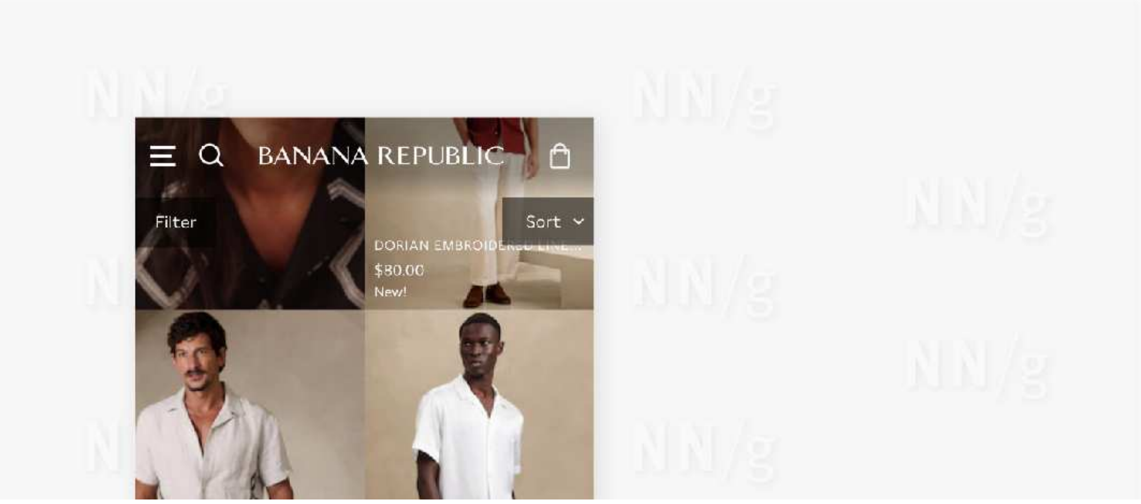
Start the evaluation by moving through the interface as if you're a user trying to complete a task. If you aren't already familiar with this product, **go through the task once just to learn** the system, without attempting to evaluate anything.

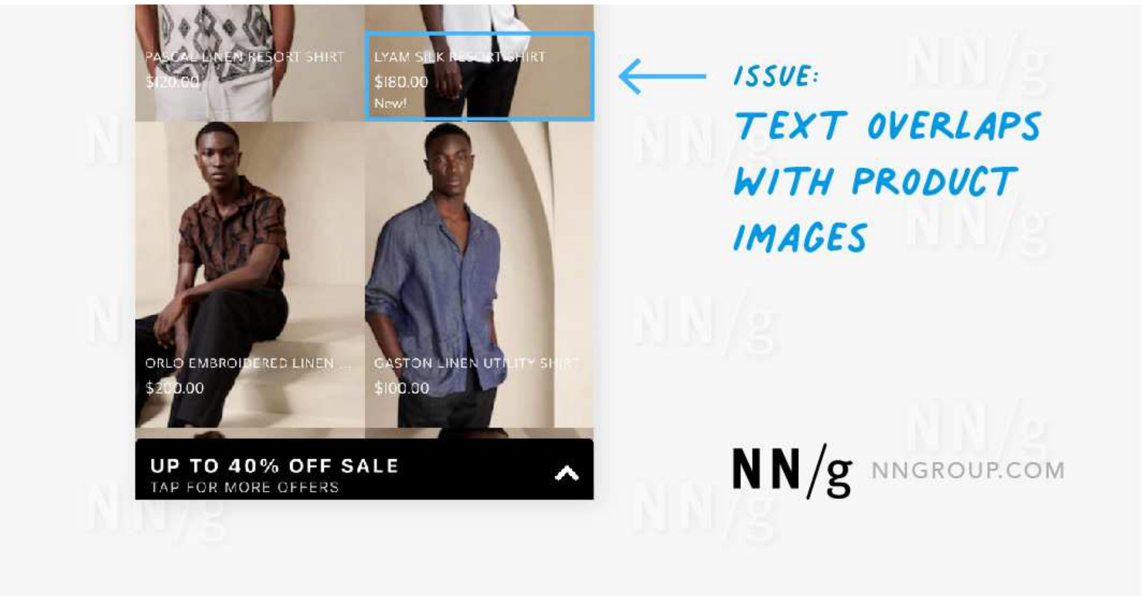
Look for Issues

Once you feel comfortable and familiar with the product, go back through the task a second time. In this second pass, look for **design elements, features, or decisions that violate one of the 10 heuristics** — in other words, they don't achieve that goal or follow that guideline. If you're using our heuristic-evaluation template, those will be the things you write down in the appropriate *Issues* section.

For example, one of the 10 heuristics is aesthetic and minimalist design (#8). This heuristic recommends that the visual design of interfaces should direct users' attention and help them achieve their goals. The product should not feel visually overwhelming or distracting.

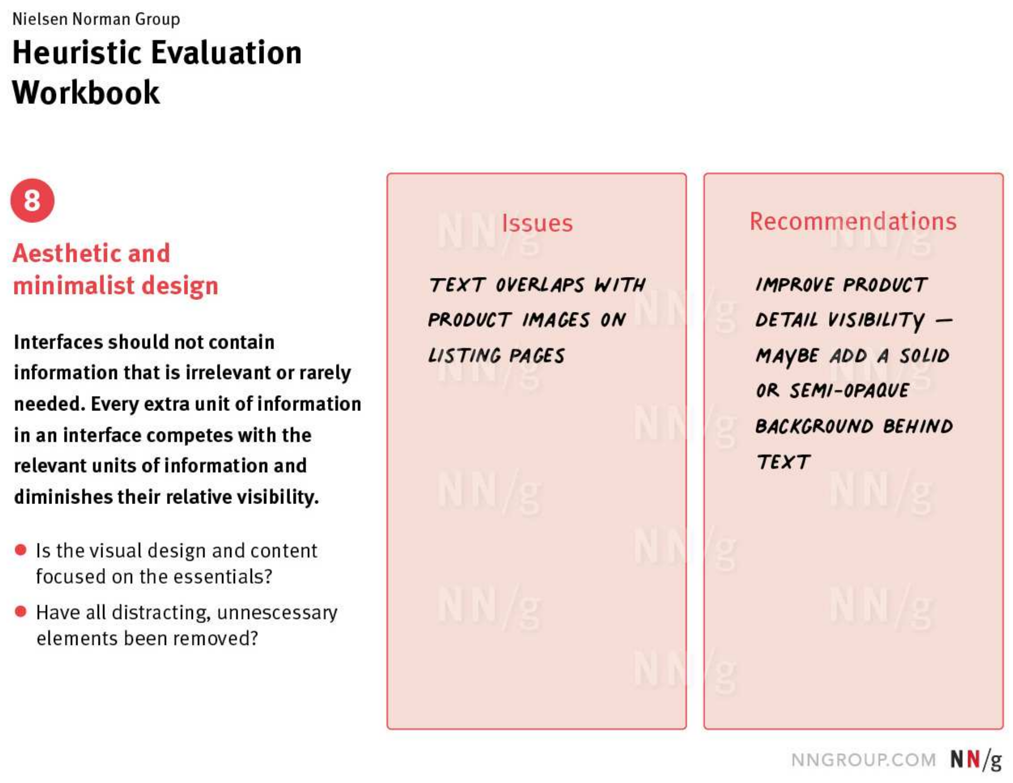
Banana Republic's listing pages layer the product details (name, price, discount) in white text directly on top of the product images. As a result, the page feels cluttered, and the text is difficult to read. This is an example of how the visual design fails to support the user's task (choosing and buying a shirt).





Banana Republic site: The text layered over images violates heuristic #8 — aesthetic and minimalist design.

In the heuristic evaluation workbook, we might write “Text overlaps with product images on listing pages” in the issues column for heuristic #8. If a recommendation for a fix comes to mind, you can note that in the second column under *Recommendations.* For example, “Improve product-detail visibility — maybe add a solid or semi-opaque background behind text.”



Nielsen Norman Group

Heuristic Evaluation Workbook

8

Aesthetic and minimalist design

Interfaces should not contain information that is irrelevant or rarely needed. Every extra unit of information in an interface competes with the relevant units of information and diminishes their relative visibility.

- Is the visual design and content focused on the essentials?
- Have all distracting, unnescessary elements been removed?

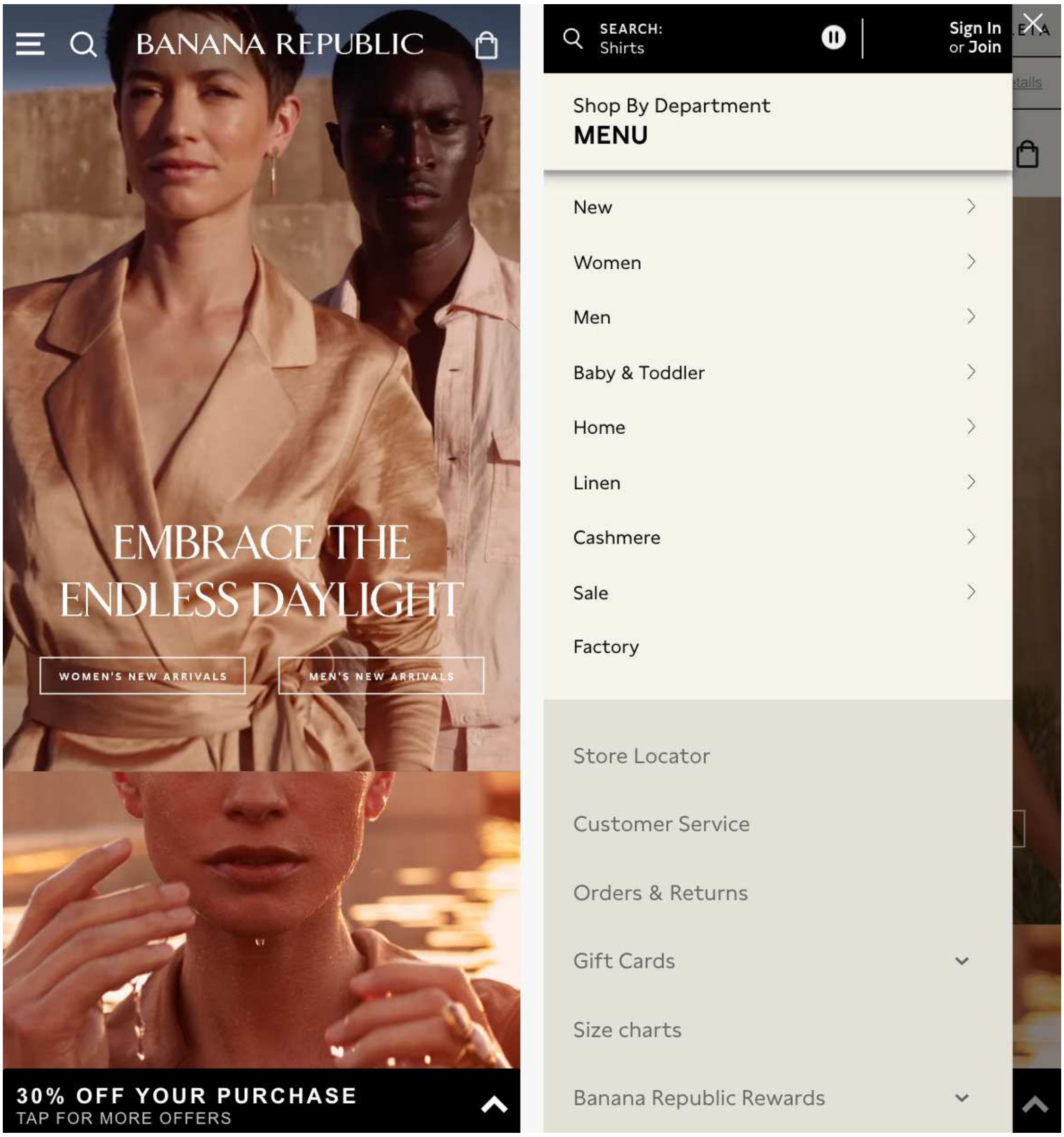
Issues	Recommendations
TEXT OVERLAPS WITH PRODUCT IMAGES ON LISTING PAGES	IMPROVE PRODUCT DETAIL VISIBILITY — MAYBE ADD A SOLID OR SEMI-OPAQUE BACKGROUND BEHIND TEXT

NNGROUP.COM NN/g

The heuristic evaluation workbook contains space to write down potential issues and recommendations for each of the 10 heuristics.

Another heuristic is recognition rather than recall (#6). As much as possible in the design, we want to reduce the burden on people's short-term memory by keeping important information visible on the screen.

We might notice that Banana Republic is using a hamburger menu as their global navigation — their navigation categories are hidden behind a menu icon in the upper left-hand corner. We might decide to document that as a violation of heuristic #6.



Hamburger menus — like the one Banana Republic's site uses — do *violate the recognition-rather-than-recall heuristic, because they hide important navigation choices, and the user has to remember to open them up to see the available options. However, that doesn't mean that a hamburger menu was the wrong design choice in this example.*

This example brings up something important to note: **just because a design choice violates a heuristic, that does not necessarily mean it's a problem** that needs to be fixed — it depends on the particular context and the available alternatives. Hamburger menus do violate heuristic #6, but, in mobile designs, that tradeoff is often necessary due to reduced screen space.

This is a great illustration of why **heuristic evaluations are not a replacement for user research.** We still need to observe our users as they are using our products to fully understand design problems.

Step 3: Consolidate Identified Issues

Once all your team members have performed their independent evaluations, it's time to synthesize the issues. Affinity diagramming (clustering similar issues) on a physical or virtual whiteboard can work well.

Discuss with your team:

- Where do we agree? Where do we disagree?
- Which issues seem most detrimental to the overall experience?
- Which issues could be most problematic for our organization or business goals?
- Which issues do we need more data on? Which should we prioritize in our next usability test?
- What steps can we take in the short and long term to address

these problems?

Exceptions

There are exceptions to these heuristics, but they're typically rare and based on context. Heuristics are guidelines, not laws, and there are some cases where you may have to violate a heuristic in pursuit of another goal (as was the case with Banana Republic's hamburger menu).

But, as Jakob Nielsen says, "you should not bet that your design is one of the few exceptions." Before deciding to intentionally violate a usability heuristic, conduct user research to ensure your rule breaking is justified.

Conclusion

Heuristic evaluations become easier the more you conduct them. With practice, you'll need to rely less and less on the actual heuristics. You'll start to develop UX instincts, so you can quickly recognize potential usability problems.

References

Nielsen, J., and Molich, R. (1990). Heuristic evaluation of user interfaces, *Proc. ACM CHI'90 Conf.* (Seattle, WA, 1-5 April), 249-256.

Nielsen, J. (1994). Heuristic evaluation. In Nielsen, J., and Mack, R.L. (Eds.), *Usability Inspection Methods*. John Wiley & Sons, New York, NY.

Nielsen, J., and Landauer, T. K. 1993. A mathematical model of the finding of usability problems. *Proceedings ACM/IFIP INTERCHI'93 Conference* (Amsterdam, The Netherlands, April

Nielsen Norman Group

Heuristic Evaluation Workbook

Use this workbook to conduct your own heuristic evaluation.

For each of Jakob's 10 Usability Heuristics, look for specific places where the interface fails to adhere to the guideline. Write your recommendations for how to fix those usability issues.

https://www.nngroup.com/articles/ten-usability-heuristics/

Nielsen Norman Group

Heuristic Evaluation Workbook

EVALUATOR:

DATE:

PRODUCT:

TASK:

1

Visibility of System Status

The design should always keep users informed about what is going on, through appropriate feedback within a reasonable amount of time.

- Does the design clearly communicate its state?
- Is feedback presented quickly after user actions?

Issues	Recommendations

2

Match Between System and the Real World

The design should speak the users' language. Use words, phrases, and concepts familiar to the user, rather than internal jargon. Follow real-world conventions, making information appear in a natural and logical order.

- Will user be familiar with the terminology used in the design?
- Do the design’s controls follow real-world conventions?

Issues	Recommendations

3

User Control and Freedom

Users often perform actions by mistake. They need a clearly marked "emergency exit" to leave the unwanted action without having to go through an extended process.

- Does the design allow users to go back a step in the process?
- Are exit links easily discoverable?
- Can users easily cancel an action?
- Is *Undo* and *Redo* supported?

Issues

Recommendations

4

Consistency and Standards

Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform and industry conventions.

- Does the design follow industry conventions?
- Are visual treatments used consistently throughout the design?

Issues

Recommendations

5

Error Prevention

Good error messages are important, but the best designs carefully prevent problems from occurring in the first place. Either eliminate error-prone conditions, or check for them and present users with a confirmation option before they commit to the action.

- Does the design prevent slips by using helpful constraints?
- Does the design warn users before they perform risky actions?

Issues	Recommendations

6

Recognition Rather Than Recall

Minimize the user's memory load by making elements, actions, and options visible. The user should not have to remember information from one part of the interface to another. Information required to use the design (e.g. field labels or menu items) should be visible or easily retrievable when needed.

- Does the design keep important information visible, so that users do not have to memorize it?
- Does the design offer help in-context?

Issues	Recommendations

7

Flexibility and Efficiency of Use

Shortcuts — hidden from novice users — may speed up the interaction for the expert user such that the design can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

- Does the design provide accelerators like keyboard shortcuts and touch gestures?
- Is content and funtionality personalized or customized for individual users?

Issues

Recommendations

8

Aesthetic and Minimalist Design

Interfaces should not contain information that is irrelevant or rarely needed. Every extra unit of information in an interface competes with the relevant units of information and diminishes their relative visibility.

- Is the visual design and content focused on the essentials?
- Have all distracting, unnescessary elements been removed?

Issues

Recommendations

9

Help Users Recognize, Diagnose, and Recover from Errors

Error messages should be expressed in plain language (no error codes), precisely indicate the problem, and constructively suggest a solution.

- Does the design use traditional error message visuals, like bold, red text?
- Does the design offer a solution that solves the error immediately?

Issues	Recommendations

10

Help and Documentation

It's best if the system doesn't need any additional explanation. However, it may be necessary to provide documentation to help users understand how to complete their tasks.

- Is help documentation easy to search?
- Is help provided in context right at the moment when the user requires it?

Issues	Recommendations

Nielsen Norman Group

Jakob’s Ten Usability Heuristics

1 Visibility *of* System Status

Designs should **keep users informed** about what is going on, through appropriate, timely feedback.

2 Match between System and the Real World

The design should speak the users' language. Use words, phrases, and concepts **familiar to the user,** rather than internal jargon.

3 User Control *and* Freedom

Users often perform actions by mistake. They **need a clearly marked "emergency exit"** to leave the unwanted state.

4 Consistency *and* Standards

Users should not have to wonder whether different words, situations, or actions mean the same thing. **Follow platform conventions.**

5 Error Prevention

Good error messages are important, but the best designs **prevent problems** from occurring in the first place.

6 Recognition Rather Than Recall

Minimize the user's memory load by making elements, actions, and options visible. Avoid making users remember information.

7 Flexibility *and* Efficiency of Use

Shortcuts — hidden from novice users — may **speed up the interaction** for the expert user.

8 Aesthetic *and* Minimalist Design

Interfaces should not contain information which is irrelevant. Every extra unit of information in an interface **competes** with the relevant units of information.

9 Recognize, Diagnose, *and* Recover from Errors

Error messages should be expressed in **plain language** (no error codes), precisely indicate the problem, and constructively suggest a solution.

10 Help *and* Documentation

It’s best if the design **doesn’t need** any additional explanation. However, it may be necessary to provide documentation to help users understand how to complete their tasks.

NN/g

www.nngroup.com/articles/ten-usability-heuristics/

HUMAN MENTAL WORKLOAD
P.A. Hancock and N. Meshkati (Editors)
Elsevier Science Publishers B.V. (North-Holland), 1988

Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research

Sandra G. Hart
Aerospace Human Factors Research Division
NASA-Ames Research Center
Moffett Field, California

Lowell E. Staveland
San Jose State University
San Jose, California

ABSTRACT

The results of a multi-year research program to identify the factors associated with variations in subjective workload within and between different types of tasks are reviewed. Subjective evaluations of 10 workload-related factors were obtained from 16 different experiments. The experimental tasks included simple cognitive and manual control tasks, complex laboratory and supervisory control tasks, and aircraft simulation. Task-, behavior-, and subject-related correlates of subjective workload experiences varied as a function of difficulty manipulations within experiments, different sources of workload between experiments, and individual differences in workload definition. A multi-dimensional rating scale is proposed in which information about the magnitude and sources of six workload-related factors are combined to derive a sensitive and reliable estimate of workload.

INTRODUCTION

This chapter describes the results of a multi-year research effort aimed at empirically isolating and defining factors that are relevant to subjective experiences of workload and to formal evaluation of workload across a variety of activities. It includes information on how people formulate opinions about workload and how they express their subjective evaluations using rating scales.

Despite much disagreement about its nature and definition, workload remains an important, practically relevant, and measurable entity. Workload assessment techniques abound; however, subjective ratings are the most commonly used method and are the criteria against which other measures are compared. In most operational environments, one of the problems encountered with the use of subjective rating scales has been high between-subject variability. We propose a rating technique by which variability is reduced. Another problem has been that the sources of workload are numerous and vary across tasks. sources of workload. The proposed rating technique, which is multidimensional, provides a method by which specific sources of workload relevant to a given task can be identified and considered in computing a global workload rating. It combines information about these factors, thereby reducing some sources of between-subject variability that are experimentally irrelevant, and emphasizing the contributions of other sources of variability that are experimentally relevant.

Conceptual Framework

We began with the assumption that workload is a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance. Thus, our definition of workload is human-centered, rather than task-centered (refs. I-12, I-22). An operator's subjective experience of workload summarizes the influences of many factors in addition to the objective demands imposed by the task. Thus, workload is not an inherent property, but rather it emerges from the interaction between the requirements of a task, the circumstances under which it is performed, and the skills, behaviors, and perceptions of the operator. Since many apparently unrelated variables may combine to create a subjective workload experience, a conceptual framework was proposed (ref. I-12) in which different sources and modifiers of workload were enumerated and related (Figure 1).

Imposed workload refers to the situation encountered by an operator. The intended demands of a task are created by its objectives, duration, and structure and by the human and system resources provided. The actual demands imposed by a task during its performance by a specific operator may be modified by a host of factors (e.g., the environment, system failures, operator errors) that are unique to that occurrence. These incidental factors may contribute either subtle or substantial sources of variability to the workload imposed by the task from one performance to the next.

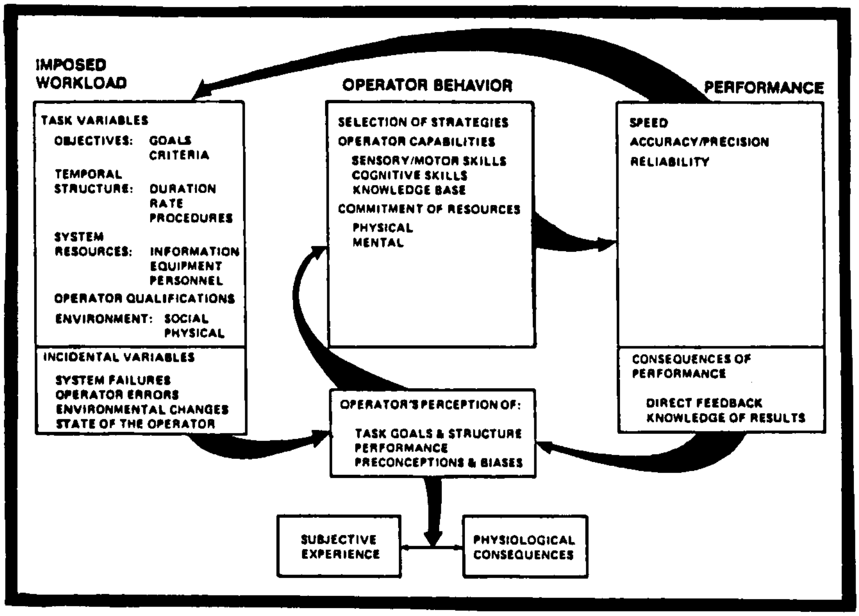


Figure 1. Conceptual framework for relating variables that influence human performance and workload.

System response refers to the behavior and accomplishments of a man-machine system. Operators are motivated and guided by the imposed demands, but their behavior also reflects their perceptions about what they are expected to do and the strategies, effort, and system resources expended to accomplish the task objectives. Operators exert effort in a variety of ways. Physical effort is the easiest to conceptualize, observe, and measure, yet its importance in advanced systems is diminishing. Mental effort serves as a potent intervening variable between measurable stimuli and measurable responses, but it is difficult to quantify directly. System performance represents the product of an operator's actions and the limitations, capabilities, and characteristics of the system controlled. Performance feedback provides operators information about their success in meeting task requirements, allowing them to adopt different strategies or exert different levels of effort to correct their own errors.

Experienced workload and physiological consequences reflect the effect on an operator of performing a task. It is the subjective experience of workload that is the legitimate domain of subjective ratings. However, it is not likely that an operator's experience of workload is a simple combination of the relevant factors. Moreover, ratings may be biased by preconceptions. Since operators are unlikely to be aware of every task variable or the processes that underlie their decisions and actions, their experiences will not reflect all relevant factors. In addition, they are influenced by preconceptions about the task and their definition of workload. Thus, we draw a distinction among the level of workload that a system designer intends to impose, the responses of a specific man-machine system to a task, and operators' subjective experiences.

The importance of subjective experiences extends beyond its association with subjective ratings. The phenomenological experiences of human operators affect subsequent behavior, and thus affect their performance and physiological responses to a situation. If operators consider the workload of a task to be excessive they may behave as though they are overloaded, even though the task demands are objectively low. They may adopt strategies appropriate for a high-workload situation (e.g., shedding tasks, responding quickly), experience psychological or physiological distress, or adopt a lower criterion for performance.

Information Provided by Subjective Ratings

In comparison with other workload assessment methods (refs. I-15, I-22), subjective ratings may come closest to tapping the essence of mental workload and provide the most generally valid and sensitive indicator. They provide the only source of information about the subjective impact of a task on operators and integrate the effects of many workload contributors. However, there are practical problems associated with translating a personal experience of workload into a formalized workload rating. People often generate evaluations about the difficulty of ongoing experiences and the impact of those experiences on their physical and mental state. However, they rarely quantify, remember, or verbalize these fleeting impressions. In fact, they may not identify their cause or effect with the concept of "workload" at all. They are aware of their current behavior and sensations and the results of cognitive processes, although they are not aware of the processes themselves (refs. I-8, I-18). Only the most recent information is directly accessible for verbal reports from short-term or working memory. Thus, a great deal of information may be available as an experience occurs; however, the experience of each moment is replaced by that of the next one. The workload of an activity may be recalled or re-created, but the evaluation is limited to whatever information was remembered, incidentally or deliberately, during the activity itself. For these and other reasons, subjective ratings do not necessarily include all of the relevant information and they may include information that is irrelevant.

Workload is experienced as a natural consequence of many daily activities. However, a formal requirement to quantify such an experience using experimentally-imposed rating scales is not a natural or commonplace activity and may result in qualitatively different responses. For this reason, Turksen and Moray (ref. I-25) suggested that the less precise "linguistic" approach provided by fuzzy logic might be appropriate for workload measurement because people naturally describe their experiences with verbal terms and modifiers (e.g., "high", "easy", or "moderate") rather than with numerical values. If workload is a meaningful construct, however, it should be possible to obtain evaluations in a variety of ways either while a task is being performed or at its conclusion.

A formal requirement to provide a rating does encourage subjects to adopt a more careful mode of evaluation, to express their judgments in a standardized format, and to adopt the evaluation criteria imposed by the experimenter. Workload evaluations are typically given with reference to arbitrary scales labeled with numbers or verbal descriptions of the magnitudes represented by extreme values. These often have no direct analog in the physical world. Since it is unlikely that individuals remember specific instances of low, medium or high workload to serve as a mental reference scale labeled "workload", absolute judgements or comparisons across different types of tasks are not generally meaningful. For features that can be measured in physical units, it is possible to distinguish among absolute, relative and value judgements from the objective information available. For workload ratings, it is relatively more difficult to distinguish between an "objective" magnitude estimate and a judgement made in comparison to an internal reference. Rating formats might include discrete numeric values, alternative descriptors, or distances marked off along a continuum. Finally, rating scales might be single-dimensional or multi-dimensional requiring judgements about several task-related or psychological variables.

Evaluating Ill-Defined Constructs

It is likely that the cognitive evaluation processes involved when people make workload assessments are similar to those adopted when they evaluate other complex phenomena. Evaluation is typically a constructive process, operating on multiple attributes of available information. It relies on a series of inferences in which the weight and value that an individual places on each piece of information may be unique and refers to their existing knowledge base (ref. I-1). Some evaluations are relatively direct, based on immediate sensory or perceptual processes, whereas others involve organization of background knowledge, inference, and relating existing knowledge to different aspects of the current situation. We feel that the experience of workload represents a combination of immediate experiences and preconceptions of the rater and is, therefore, the result of constructive cognitive processes.

In making many judgements, people apply heuristics that are natural to them and seem to be appropriate to the situation. Heuristics simplify evaluation and decision processes because they can be applied with incomplete information, reducing the parameters that must be considered by relating the current situation to similar events in the rater's repertoire. However, their use may lead to systematic biases (ref. I-26). Different components of a complex construct may be particularly salient for one individual but not for another and for one situation but not another. Thus, different information and rules-of-thumb may be considered.

The heuristics used to generate evaluations of various physical features can be determined systematically. This is done by varying different features of an object and comparing the evaluations to the objective magnitudes of the components. If there is a direct mapping between an increase in a relevant physical dimension and the obtained evaluation, the nature of the relationship can be identified. These relationships are not likely to be linear, however. Rather, noticeable differences in one or more dimensions are proportional to the magnitude of the change. In addition, by varying the wording of written or verbal instructions, or presenting different reference objects, the basis and magnitude of judgements can be manipulated (ref. I-10, I-11).

When people evaluate the workload of a task there is no objective standard (e.g., its "actual" workload) against which their evaluations can be compared. In addition there are no physical units of measurement that are appropriate for quantifying workload or many of its component attributes. This absence of external validation represents one of the most difficult problems encountered in evaluating a candidate workload assessment technique or the accuracy of a particular rating. There is no objective workload continuum, the "zero" point and upper limits are unclear, and intervals are often arbitrarily assigned. The problem of a "just noticeable difference" is particularly acute in workload assessment, since rating dimensions are often indirectly related to objective, quantifiable, physical dimensions.

The attributes that contribute to workload experiences vary between tasks and between raters because workload is not uniquely defined by the objective qualities of the task demands; workload ratings also reflect an operator's response to the task. Thus, the workload experiences of different individuals faced with identical task requirements may be quite different because the relationship between objective changes in a task and the magnitudes of workload ratings is indirect rather than direct. This factor distinguishes workload ratings from many other types of judgements. Furthermore, if workload is caused by one particularly salient source or by very high levels of one or more factors, then it is likely that other factors will not be considered in formulating a workload judgement. Specific workload-related dimensions might be so imperative, or so imbedded in a particular context, that they contaminate other, less subjectively salient factors. Conversely, less salient factors cannot be evaluated without also considering those that are more salient.

Individuals' Workload Definitions

Two facets of subjective workload experiences are of interest: the immediate, often unverbalized impressions that occur spontaneously, and a rating produced in response to an experimental requirement. It is unlikely that the range of ratings that subjects typically give for the same task reflects misinterpretation of the question--most people have some concept of what the term workload means. However, they use the most natural way to think about it for themselves. Individuals may consider different sets of variables, (which may be identical to those experimenter intended) because they define (and thus experience) workload in different ways. The amount of "work" that is "loaded" on them, the time pressure under which a task is performed, the level of effort exerted, success in meeting task requirements, or the psychological and physiological consequences of the task represent the most typical definitions. Thus, one individual's "workload" rating may reflect her assessment of task difficulty while another's might reflect the level of effort he exerted. It is impossible to identify the source or sources of a workload rating from the magnitude of the numeric value.

In general, people are unaware of the fuzziness of their own definitions or the possibility that theirs might be different than someone else's. Given more information about what factors they should consider, they can evaluate these factors (e.g., they can rate stress, fatigue, frustration, task demands, or effort) even though they might not naturally include them in a subjective experience of workload. However, it seems to be intuitively unlikely that their global, personal experiences of workload would be affected by instruction to consider only one or two aspects of a situation.

Thus, we assume that workload represents a collection of attributes that may or may not be relevant in controlling assessments and behavior. They depend on the circumstances and design of a given task and the *a priori* bias of the operator. The natural inclinations of different individuals to focus on one task feature or another may be overwhelmed by the types and magnitudes of factors that contribute to the workload of a specific task. For example, the workload of one task might be created by time pressure, while that of another might be created by the stressful conditions under which it was performed. The workload of each task can be evaluated, but the two apparently comparable ratings would actually represent two different underlying phenomena.

Sources of Rating Variability

Workload ratings are subject to a variety of task- and operator-specific sources of variability, some of which have been mentioned above (e.g., identifiable biases held by the raters or the objective manipulations of task parameters). Others represent the less predictable, but measurable, behavioral responses of operators to the task. The remainder are more difficult to identify: differences in sensitivity to the types and magnitudes of task manipulations, motivation, expectations, and subjective anchor points and interval values. The large between-subject variability characteristic of subjective ratings does not, therefore, occur exclusively as a consequence of random error or "noise". Instead, many of the sources of variability can be identified and minimized through giving instructions, calibrating raters by demonstrating concrete examples, providing reference tasks, and identifying subjective biases and natural inference rules. The workload experiences of operators are difficult to modify, but the procedures with which evaluations are obtained can b(designed to reduce unwanted between-subject sources of variability.

Research Approach

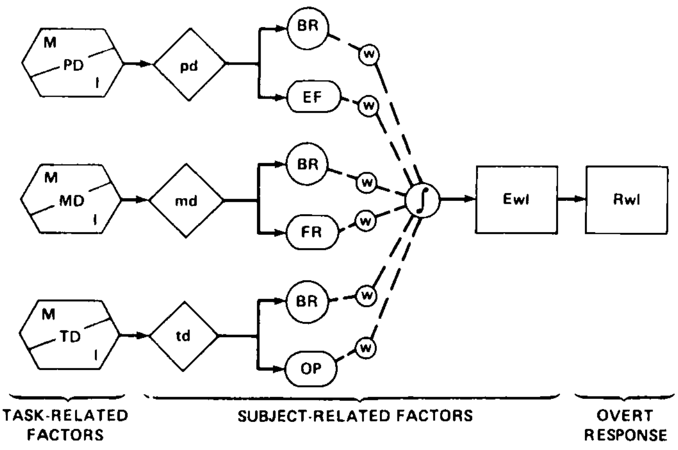
The goal of the research described below was to develop a workload rating scale that provides a sensitive summary of workload variations within and between tasks that is diagnostic with respect to the sources of workload and relatively insensitive to individual differences among subjects. We formulated a conceptual framework for discussing workload that was based on the following assumptions: workload is a hypothetical construct; it represents the cost incurred by human operators to achieve a specific level of performance and is not, therefore, uniquely defined by the objective task demands; and it reflects multiple attributes that may have different relevance for different individuals; it is an implicit combination of factors. Although the experience of workload may be commonplace, the experimental requirement to quantify such an experience is not. Nevertheless, subjective ratings may come closest to tapping the essence of mental workload and provide the most generally valid, sensitive and practically useful indicator. The ability of subjects to provide numerical ratings has received limited theoretical attention because ratings are subject to "undesirable" biases. In fact, these biases may reflect interesting and significant cognitive processes (ref. I-1). In addition, although there may be wide disagreement among subjects in the absolute values of ratings given for a particular task, the rank-ordering of tasks with respect to workload is quite consistent and the magnitudes of differences in ratings among tasks are reasonably consistent. There is a common thread that unites subjective ratings that can be termed "workload". The problem is how to maximize the contribution of this unifying component to subjective ratings, and to identify and minimize the influences of other, experimentally irrelevant, sources of variability.

To accomplish this, a set of workload related factors was selected and subjective ratings were obtained in order to determine the following: (1) What factors contribute to workload? (2) What are their ranges, anchor points, and interval values? (3) What subset of these factors contributes to the workload imposed by specific tasks? and (4) What do individual subjects take into account when experiencing and rating workload? The following sections review the results of a series of experiments that were undertaken to provide such a data base. The goal was to provide empirical evidence about which factors individuals do, or do not associate with the experience of workload and the rules by which these factors are combined to generate ratings of overall workload.

First, we analyzed the data within each experiment to determine the sensitivity of individual scales, overall workload (OW) ratings, and weighted workload (WWL) scores to experimental manipulations. Next, the data from similar experiments were merged into six categories. Correlational and regression analyses were performed on these data, as well as on the entire data base, to determine (1) the statistical association among ratings and (2) the degree to which these scales, taken as a group, predicted OW ratings. The results of these analyses were then used to select a limited set of subscales and the weighting procedure for a new multi-dimensional workload rating technique.

We found that, although the factors that contributed to the workload definitions of individual subjects varied as predicted, task-related sources of variability were better predictors of global workload experiences than subjective biases. A model of the psychological structure of the subjective workload estimation process evolved from the analyses performed on this data base. It is presented in Figure 2.

This model represents the psychological structure of subjective workload evaluations. It is adapted from a similar structure proposed by Anderson (ref. I-1) for stimulus integration, since the process of workload assessment is almost certainly an integrative process in which external events are translated into subjective experiences and overt responses. The objective mental, physical, and temporal demands (MD.PD and TD) that are imposed by a task are multi-dimensional and may or may not covary. They are characterized by objective magnitudes (M) and levels of importance (I) specific to a task. When the requirements of a task are perceived by the performer, their significance, magnitudes, and meaning may be modified somewhat depending on his level of experience, expectations, and understanding. These psychological variables, which are counterparts to the objective task variables, are represented by md, pd, and td. They yield emotional (e.g., FR), cognitive, and physical (e.g., EF)



PD, MD, TD	Objective physical, mental and temporal task demands
M, I	Objective magnitudes and importance of sources of demands
pd, md, td	Psychological representations of task demands
BR	Behavioral responses to task demands
OP, EF, FR	Subjective responses/evaluations of behavioral responses
w	Subjective weighting of factors
Ewl	Integrated subjective experience of workload
Rwl	Formal numeric or verbal evaluation of workload

Figure 2. A model of the subjective workload estimation process.

responses that may be evidenced as measurable overt behaviors (BR). The results of the individuals' actions may be self-evaluated (e.g., OP), thereby leading to adjustments in the levels or types of responses or a re-evaluation of task requirements. These subjective evaluations, too, may or may not covary with each other and, although they are related to the objective demands, specific stimulus attributes may differentially influence behavior under different circumstances. Subjectively weighted (w) combinations of such variables can be integrated into a composite experience of workload (Ewl). This implicit experience may be converted into an explicit workload rating (Rwl) in response to an experimental requirement. The resulting values do not represent inherent properties of the objective demands. Rather, they emerge from their interaction with a specific operator. In order to predict and understand the relationship between objective task manipulations and rated workload, the salient factors and the rules by which they are objectively and subjectively combined must be identified and an appropriate procedure developed to obtain an accurate summary evaluation.

Thus, two types of information are needed about each factor included in a multidimensional workload scale: (1) its subjective importance as a source of loading for that type of task (its weight), and (2) its magnitude in a particular example of the task (the numerical value of a rating). For example, the mental demands of a task can be the most salient feature of its demand structure, although the amount of such demands can vary from one version of the task to another. Conversely, the value of one might vary at different levels of the other: time pressure might become relevant only when it is high enough to interfere with performance.

A rating scale is proposed, the NASA-Task Load Index (NASA-TLX), that consists of six component scales. An average of these six scales, weighted to reflect the contribution of each factor to the workload of a specific activity from the perspective of the rater, is proposed as an integrated measure of overall workload. Finally, the results of a validation and reliability study are described. See Reference Section III for a listing of recent experimental uses of the NASA-TLX.

Research Objectives and Background

Our first step was to ask people engaged in a wide range of occupations to identify which of 19 factors were subjectively equivalent to workload, related to it, or unrelated (ref. I-13). Surprisingly, none of the factors was considered to be irrelevant by more than a few raters, and at least 14 of the factors were considered to be subjectively equivalent to workload by more than 60% of them. No relationship between the response patterns and the evaluators' educational or occupational backgrounds were found.

Our next step was to ask several groups of subjects to evaluate their experiences with respect to the 14 most salient factors following a variety of laboratory and simulated flight tasks (refs. I-2, I-14,I-29). Different concepts of workload were identified by determining which component ratings covaried with an overall workload rating that was provided by each subject after each experimental condition. Several factors (e.g., task difficulty and complexity, stress, and mental effort) were consistently related to workload across subjects and experiments. Other factors (e.g., time pressure, fatigue, physical effort, and own performance) were closely related under some experimental conditions, and not under others.

Again, the most salient factors were selected and a set of 10 bipolar rating scales were developed (Figure 3): Overall Workload (OW), Task Difficulty (TD), Time Pressure (TP), Own Performance (OP), Physical Effort (PE), Mental Effort (ME), Frustration (FR), Stress (ST), Fatigue (FA), and Activity Type (AT). AT represented the levels of behaviors identified by Rasmussen (ref. I-19): skill-based, rule-based, and knowledge-based. It has been suggested that the three levels of behavior are associated with increasing levels of workload (refs. I-16, I-

FIGURE 3: RATING SCALE DESCRIPTIONS

Title	Endpoints	Descriptions
OVERALL WORKLOAD	*Low, High*	The total workload associated with the task, considering all sources and components.
TASK DIFFICULTY	*Low, High*	Whether the task was easy or demanding, simple or complex, exacting or forgiving.
TIME PRESSURE	*None, Rushed*	The amount of pressure you felt due to the rate at which the task elements occured. Was the task slow and leisurely or rapid and frantic?
PERFORMANCE	*Failure, Perfect*	How successful you think you were in doing what we asked you to do and how satisfied you were with what you accomplished.
MENTAL/SENSORY EFFORT	*None, Impossible*	The amount of mental and/or perceptual activity that was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.).
PHYSICAL EFFORT	*None, Impossible*	The amount of physical activity that was required (e.g., pushing, pulling, turning controlling, activating, etc.).
FRUSTRATION LEVEL	*Fulfilled, Exasperated*	How insecure, discouraged, irritated, and annoyed versus secure, gratified, content, and complacent you felt.
STRESS LEVEL	*Relaxed, Tense*	How anxious, worried, uptight, and harresed or calm, tranquil, placid, and relaxed you felt.
FATIGUE	*Exhausted, Alert*	How tired, weary, worn out, and exhausted or fresh, vigorous, and energetic you felt.
ACTIVITY TYPE	*Skill Based, Rule Based, Knowledge Based*	The degree to which the task required mindless reaction to well-learned routines or required the application of known rules or required problem solving and decision making.

28). Each scale was presented as an 12-cm line with a title (e.g., MENTAL EFFORT) and bipolar descriptors at each end (e.g., HIGH/LOW). Numerical values were not displayed, but values ranging from 1 to 100 were assigned to scale positions during data analysis. This set of scales was used to evaluate the experiences of subjects in 25 different studies. The ratings were obtained after each experimental task. The results obtained in 16 of these experiments are the focus of the current chapter. Since the research questions and environments differed from one experiment to the next, the data base includes a broad set of experiences in which the associations among workload-related factors, global ratings of workload, and measures of performance could be evaluated.

The relative importance of the nine component factors to each subject's personal definition of workload was determined in a pretest. All possible pairs (n = 36) of the nine factors were presented in a different random order to each subject. The member of each pair selected as most relevant to workload was recorded and the number of times each factor was selected was computed. The resulting values could range from 0 (not relevant) to 8 (more important than any other factor). The more important a factor was considered to be, the more weight the ratings of that factor were given in computing an average weighted workload score (WWL) for each experimental condition. These data were obtained for two reasons: (1) to examine the relationship between the expressed biases of subjects about each factor and the associations between the magnitude of the ratings for the same factors and rated OW, and (2) to use these as weights in combining the nine bipolar ratings to produce a workload score that emulated the heuristics that subjects reported using.

In computing the weighted workload scores, we assumed the following: (1) The factors considered in formulating a single OW rating varied from one subject to the next, contributing to between-subject (B-S) variability. (2) Subjects would be able to evaluate all of the factors (even though they might not normally consider them in evaluating workload). (3) The subjects could judge the magnitudes of the component factors more accurately and with less B-S variability than they could the fuzzier concept of OW. (4) The ratings the subjects made might represent the "raw data" for subjects' natural inference rules. (5) By combining these component judgements according to each subject's own inference rules (as reflected in the workload weights), an estimate of workload could be derived (WWL) that would be less variable between subjects than ratings of OW. (6) The combination rules would be linear. (7) The weighted averaged ratings would reflect the general importance of the factors to individual subjects and their rated magnitudes in a given task.

Our goal was to determine which scales best reflected experimental manipulations within experiments, differentiated among different types of activities, provided independent information, and were subjectively and empirically associated with global workload ratings. To accomplish this, we attempted to obtain information about the individual and joint relationships among the nine factors, OW, and experimental manipulations from many perspectives to obtain the most complete understanding of the underlying functions.

OVERALL RESULTS

The experiments included in the data base described in this chapter are listed in Reference Section II. Each one was analyzed individually and the relationships among performance measures, ratings, WWL scores, and experimental variables have been reported elsewhere. Thus, specific experimental results will not be described below. Instead, more global statements germane to the definition and evaluation of workload in general will be made for categories of similar experiments and the entire data base. Although many of the same subscales and the weighting technique were used in other experiments, these were not included either because the raw data were not readily available or because one or more subscales were not used (refs. I-5, I-17, I-27, I-28).

The data were divided into two "population" data bases. The rating data base contained 3461 entries for each of the 10 scales and WWL. The weight data base contained the workload biases given by the same 247 subjects. Figure 4 presents the average weights given to the nine factors. and presents the average ratings. Tables 1a and 1b show the correlations among the weights placed on each factor and among the ratings, respectively. Figure 5 presents the relative frequency distributions of obtained ratings and WWL scores.

A variety of statistical analyses were performed within individual experiments to demonstrate the effectiveness of the experimental manipulations. They included analyses of variance and correlations among measures of workload and performance. In addition, multiple correlations among individual rating scales were performed, the coefficients of variation (SD/Mean) for OW and for WWL were computed for individual experimental conditions, and sensitivity tests were conducted to compare the percentages of variance accounted for by the OW rating scale and the WWL score. Additional analyses were also performed on the groups of data in each category and for the entire data base. Non-parametric Komalgorov-Schmirnoff tests (ref. I-23) were performed to compare distributions of ratings given for each scale among the categories of experiments and against the "population" data base. Standard multiple correlations were performed among the scales and among the workload-importance weights.

The individual scales were correlated with OW to determine the associations of each one with the more global construct across all categories and within each category. In addition, all nine scales were regressed against OW to determine the percent of variance in OW ratings for which their linear combination accounted.

Stimulus attributes were under only limited experimental control and may have been too inter-correlated to discriminate among the range of individual dimensions represented in either individual or collective experiments. Furthermore, the variability in generating workload ratings may not have depended solely on the experimentally imposed tasks (ref. I-1) because raters may or may not have perceived the task parameters in the same way (which could lead to a subject by task interaction). Finally, the fact that there was multi-collinearity among the component scales suggests that the beta weights for individual factors may not have reflected their individual and joint predictive power. Nevertheless, the beta weights (Table 2a) taken in conjunction with the correlations between each factor and OW enabled us to identify the primary sources of workload in each type of task. For simplicity's sake, any correlation that accounted for more than 50 percent of the variance will be considered. The squared correlation coefficients for each factor with OW are presented in Table 2b.

Weights

Although there was considerable disagreement among subjects about which combinations of factors best represented their concept of workload, some consistent trends were observed (Figure 4a). TP was considered the most important variable, followed by FR. ST, ME and TD. PE was considered the least important variable and FA and AT were also relatively unimportant. The importance assigned to each factor appeared to be relatively independent of that assigned to any other (Table 1a). To some extent this is an artifact of the pairwise comparison technique with which the weights were obtained; every decision in favor of one member of a pair of factors was made at the expense of whatever factor was not selected. The greatest statistical association was found between AT and ST (-0.50) or FR (-0.40); if the type of activity performed was considered particularly important, feelings of ST or FR were not considered relevant, and *vice versa*. The next highest degree of association was found between OP and FA (-0.46) or ST (-0.35); subjects who equated workload with success or failure on a task did not consider their feelings of FA or ST to be relevant and *vice versa*. This suggests that there may be at least two patterns of workload definition: one based on task and

Table 1a: POPULATION

Correlations among subjective importance values of 9 workload-related factors

	TD	TP	OP	PE	ME	FR	ST	FA
TP	.05							
OP	-.08	-.24						
PE	-.12	-.31	-.07					
ME	.16	-.24	-.01	-.05				
FR	-.37	.05	-.21	-.26	-.30			
ST	-.21	.07	-.24	-.35	-.28	.32		
FA	-.21	-.03	-.46	.03	-.36	.10	.24	
AT	.08	-.17	.08	.17	.30	-.40	-.50	-.34

Table 1b: POPULATION

Correlations among raw bipolar ratings and OW

	TD	TP	OP	PE	ME	FR	ST	FA	AT
TP	.64								
OP	.58	.50							
PE	.53	.57	.38						
ME	.76	.58	.53	.47					
FR	.65	.60	.68	.45	.61				
ST	.63	.66	.48	.56	.60	.71			
FA	.38	.33	.40	.40	.37	.51	.52		
AT	.28	.29	.11	.20	.30	.21	.21	.11	
OW	.83	.60	.50	.52	.73	.63	.62	.40	.30

Table 2a

Beta weights for ratings regressed on OW (*=p<.01)

	r^2	TD	TP	OP	PE	ME	FR	ST	FA	AT
SINGLE-COGNITIVE	.75	.50*	.02	.13*	.06	.16	-.03	.09*	.07*	.06
SINGLE-MANUAL	.81	.47	*.13*	-.14*	.11*	.28*	-.02	.26*	-.03	-.02
DUAL-TASK	.85	.49*	.11*	-.11*	.13*	.34*	.01	.03	.10*	-.01
FITTSBERG	.80	.56*	.03	.05	.04	.18*	.04	.10*	.02	.06
POPCORN	.65	.48*	.23*	-.12*	.02	-.07*	.17*	.09*	-.08*	.07*
SIMULATIONS	.77	.79*	.03	.05	.04	.22*	-.10*	.05	-.10*	.09*
POPULATION	.73	.55*	.09*	-.02	.07*	.21*	.01	.10*	-.01	.01

Table 2b

Variance in OW accounted for by each factor for each experimental category

	TD	TP	OP	PE	ME	FR	ST	FA	AT
SINGLE-COGNITIVE	.69	.26	.25	.14	.52	.41	.30	.17	.14
SINGLE-MANUAL	.69	.36	.19	.26	.58	.48	.52	.20	.05
DUAL-TASK	.77	.58	.34	.36	.71	.49	.50	.19	.18
FITTSBERG	.74	.44	.15	.26	.58	.48	.38	.18	.16
POPCORN	.59	.55	.29	.19	.40	.37	.37	.09	.09
SIMULATIONS	.74	.13	.14	.18	.42	.11	.20	.04	.01
POPULATION	.69	.36	.25	.27	.53	.39	.38	.16	.09

performance related factors and another based on the subjective and physiological impact of tasks on the performer.

Ratings

The grand means of the 10 scales across all of the experiments were not equivalent (Figure 4b). This suggests either that the range of tasks was not sufficiently representative of the possible ranges for different scales, or that the bipolar descriptions used to anchor the scales were not subjectively equivalent. Average ratings given for the 10 scales ranged from 25 (PE) to 42 (ME). Overall rating variability was relatively consistent across the ten scales (SDs ranged from 20 to 24). As expected, the WWL scores were less variable (SD = 17).

Figure 5 depicts the frequency distributions of ratings obtained across all experiments and subjects for each factor. The relative frequencies represent the average magnitude of ratings on each factor scaled in 10 point increments. The distributions of individual scales were quite different. TD, OP, ME, and OW ratings, and WWL scores were normally distributed across subjects and experiments. TP, ST, FA, and PE distributions were skewed; most of the ratings were relatively low, but there were instances in which very high values were given. AT ratings were bimodally distributed. The peaks centered between the points designated "skill-based" and "rule-based" and between those designated as "rule-based" and "knowledge-based". Each distribution was compared to every other using the Komalgorov-Schmirnoff test. Significant differences were found among all of the distributions except among OW, TD, and TP. The greatest differences were found between WWL scores (which combines elements from all of the other scales weighted to reflect the individual subject's biases) and the individual scales.

The rank-order correlation between mean OW ratings and WWL scores within each experiment and across all experiments was very high (0.99). However, the coefficients of variation were substantially less for the WWL scores (0.39) than for OW ratings (0.48). Thus, the reduction in variability found for WWL scores was not simply due to the smaller magnitudes of these scores (mean = 35) compared to OW ratings (mean = 39) but represented a meaningful reduction of unwanted "noise". Thus, the linear combination of ratings, weighted according to the information available about each subject's natural inference rules, discriminated among experimental conditions at least as well as a single OW rating. More significant,

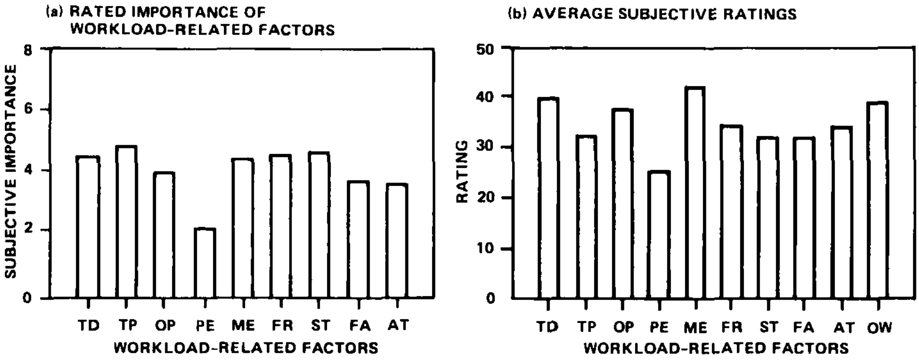


Figure 4. Summary of *a priori* importance (4a) and task-related magnitudes (4b) assigned to ten factors by all subjects (Ns = 247) and for all experimental conditions (Ns X Nc = 3461).

however, was the finding that B-S variability was less for WWL scores than for OW ratings in every experiment. The coefficients of variation were computed for each experimental condition and averaged for each experiment. They ranged from 0.19 to 0.73 for OW ratings and from 0.17 to 0.60 for WWL scores. The average reduction in variability was 20% between OW ratings and WWL scores, although it was as great as 46% for some experiments. Also, in all cases, differentially weighting the bipolars to produce WWL reduced B-S variability and increased sensitivity to experimental manipulations beyond that which could be obtained by computing a simple average of individual scales. The B-S variability of the equal weighting scheme fell between that of WWL and the OW ratings. Thus, we were able to synthesize a workload estimate from the elemental values given by the subjects (the bipolar ratings) by combining them according to an approximation of their own inference rules (the weights). This derived score appeared to reflect a common factor in each experimental condition (its overall workload), but with less variability among subjects than OW ratings.

A significant, positive association was found among many of the rating scales (Table 1b). Most of the correlations were significant, because so many data points were included, but not all of them accounted for a meaningful percentage of variance. The highest correlations were found between ME and TD (0.76) and between ST and FR (0.71); however, only the correlations between TD and OW and between ME and OW accounted for more than 50 percent of the variance (Table 2b).

TD, ME, and ST had the highest loadings in the regression equation that related ratings on the nine component factors to OW (0.55, 0.21, and 0.10, respectively) (Table 2a). Although FR was significantly correlated with OW, it contributed nothing to the OW regression equation. This could reflect the fact that it was so highly correlated with most of the other factors (e.g., TD, TP, OP, ME, ST, FA) that it did not contribute independently to OW. TP, often considered to be a primary component of workload, contributed surprisingly little to the regression equation (loading = 0.09). It is possible that this occurred because TP was not deliberately manipulated as a source of loading in many of the experiments. AT was notably unrelated to the other factors and did not contribute significantly to the OW regression equation. FA, also, was relatively unrelated to the other scales, most likely because the effects of fatigue were counterbalanced across experimental conditions (by varying the order of presentation for different levels) in most of the studies.

Table 3

A priori rank-order of factors (weights) compared to empirical associations with OW ratings

	Weight	Loading	Correlation with: OW
TP	4.75	.09	.60
TD	4.50	.55	.83
ME	4.36	.21	.73
OP	3.95	-.02	.50
ST	4.56	.10	.62
FR	4.51	.01	.63
FA	3.56	-.01	.40
AT	3.60	.01	.30
PE	2.21	.07	.52

It is interesting to compare the associations between the nine factors and workload as expressed in the preliminary pairwise comparisons to the empirical relationships observed between ratings on the same factors and OW ratings. Table 3 summarizes the *a priori* evaluations (the weights), the loadings for each factor in the OW regression equation, and the correlations between ratings on each scale and OW ratings across all subjects and experimental conditions. As you can see, there were some discrepancies. Most notably, TP was judged to be more closely related to OW (it was given the highest weight) than was apparent from the experimental results. The same was true for OP. On the other hand, PE was rarely selected as an important component of workload (it was given the lowest weight), but ranked 5th in the regression equation. These results, taken in combination with the success of the derived workload score in reducing B-S variability without substantially improving sensitivity to experimental manipulations, suggest that other factors influenced the association between component factors and OW in addition to the differences among subjects' workload definitions.

EXPERIMENTAL CATEGORIES

The data from similar types of tasks were grouped into six categories to determine whether different sources of loading (e.g., mental or physical effort, time pressure, task difficulty) did in fact contribute to the workload of different kinds of activities. Some studies

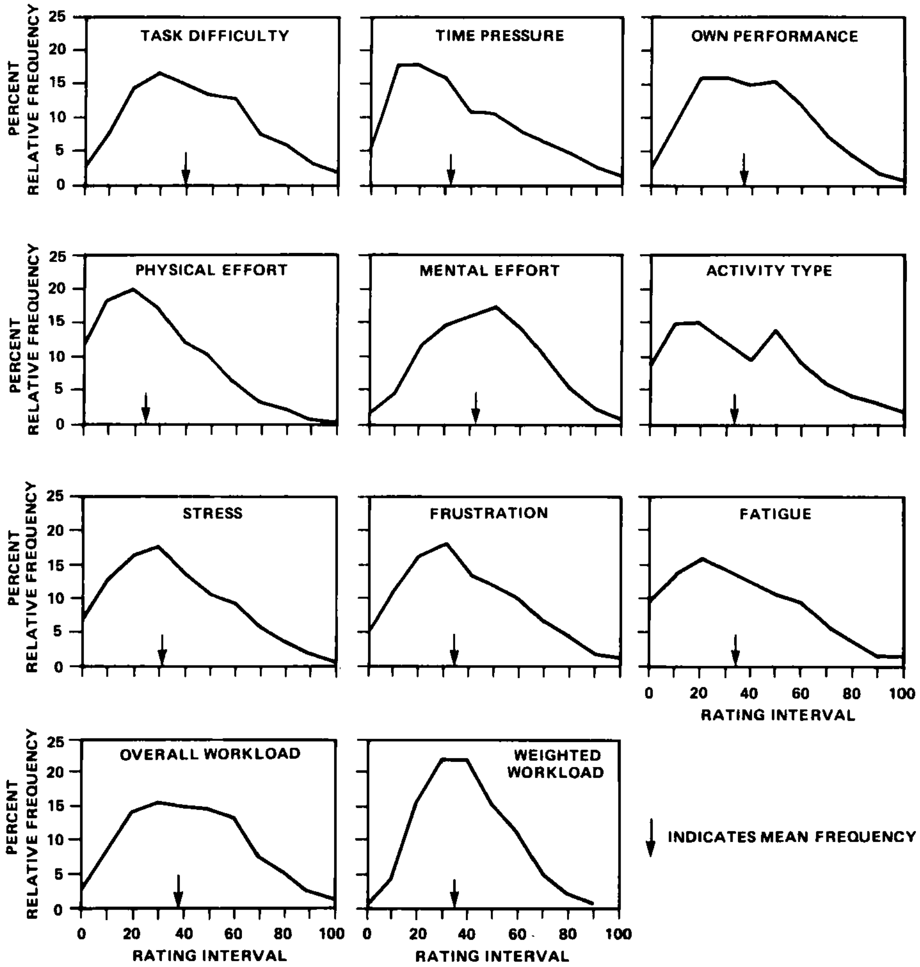


FIGURE 5. RELATIVE FREQUENCY DISTRIBUTIONS OF RATINGS AND WWL SCORES FOR ALL SUBJECTS AND EXPERIMENTAL CONDITIONS (Nc X Ns = 3461).

provided data from different experimental conditions for more than one category. The categories are

(1) Simple, discrete tasks that emphasized **SINGLE COGNITIVE** activities (refs. II-2, 6, 7, 10, 11, 13, 14),

(2) Continuous **SINGLE**-axis **MANUAL** control tasks (refs. II-2, 14),

(3) **DUAL-TASK** experiments pairing concurrent but unrelated cognitive and manual control activities (refs. II-2, 15),

(4) **FITTSBERG** tasks where response selection and execution elements were functionally integrated and sequentially executed (refs. II-6, 7, 11, 13, 16),

(5) **POPCORN** task supervisory control simulations (refs. II-1, 4, 5),

(6) **SIMULATIONS** conducted in a motion-base, single-pilot, simulator (refs. II-3, 8, 19).

The same analyses that were performed on the "population" data bases were performed for each experimental category. In addition, each category was compared to the "population". The presence of task-related sources of variability in workload was determined by examining the correlation matrices of factors, the correlation tables of factors by categories, and the regressions of the subscales on OW (Table 2a).

Our expectation was that different factors would contribute in different amounts to the overall workload of various types of tasks. For example, ME should be more salient for the SINGLE-COGNITIVE tasks, whereas PE should be more important for the SINGLE-MANUAL tasks. TP should be a particularly important source of workload for the POPCORN tasks, as this was the primary factor that was experimentally manipulated, whereas it should play a minor role in the FITTSBERG tasks, as TP was not deliberately manipulated there.

We assumed that the subjects included in each category represented a random sampling from the population as a whole and that there would be no systematic differences in workload biases of subjects who participated in one category of experimental tasks as compared to another. Since the workload biases were obtained in advance of each experiment, they should represent relatively stable opinions held by the subjects, rather than the effects of specific experimental manipulations. In fact, this was what we found. However, considerable variability was expected within each category due to the individual differences that are the focus of the weighting technique. Because the weights given by the subjects in each category were not significantly different from the population, the specific values obtained for each category will not be presented.

SINGLE-COGNITIVE Category

The SINGLE-COGNITIVE category included data from seven experiments. Each experimental task generally presented one stimulus and required one response for each trial. The primary source of loading was on cognitive processes. Five groups of experimental conditions were the single-task baseline levels for other experiments. The tasks included (1) a spatial transformation task presented visually or auditorily and performed vocally or manually; (2) variants of the Sternberg memory search task presented visually or auditorily; (3) choice reaction time; (4) same/different judgements; (5) mental arithmetic; (6) time estimation; (7) greater/less than judgements; (8) entering a number or a number plus a constant with

FIGURE 6A. SINGLE-COGNITIVE CATEGORY: SUMMARY OF RATINGS (Ns X Nc = 554).

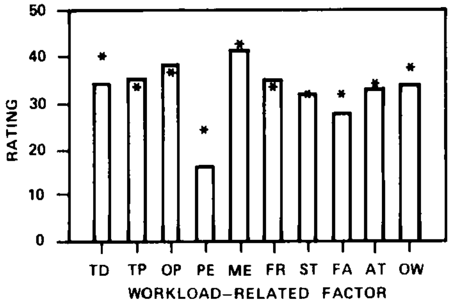


FIGURE 6B. SINGLE-MANUAL CATEGORY: SUMMARY OF RATINGS (Ns X Nc = 240).

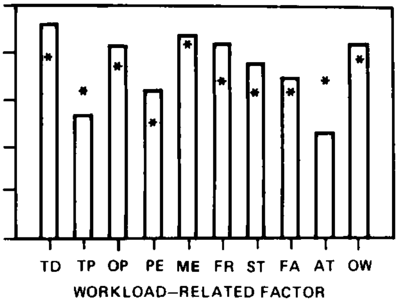


FIGURE 6C. DUAL-TASK CATEGORY: SUMMARY OF RATINGS (Ns X Nc = 732).

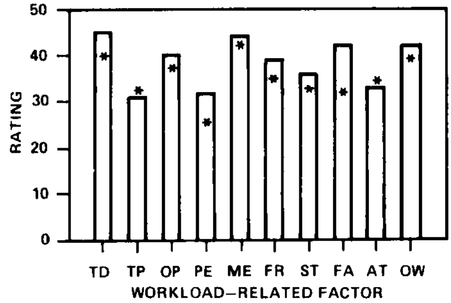


FIGURE 6D. FITTSBERG CATEGORY: SUMMARY OF RATINGS (Ns X Nc = 918).

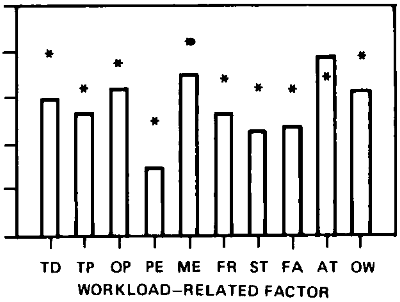


FIGURE 6E. POPCORN CATEGORY: SUMMARY OF RATINGS (Ns X Nc = 504).

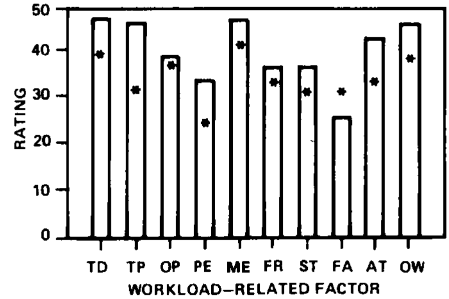
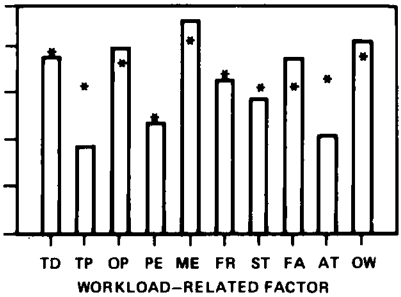


FIGURE 6F. SIMULATION CATEGORY: SUMMARY OF RATINGS (Ns X Nc = 396).



*INDICATES GRAND MEAN OF POPULATION (N = 3461) different input devices; (9) memory span; (10) flight-related heading calculations; and (11) mental rotation.

Performance was evaluated by percent correct and reaction time (RT). The typical finding was that accuracy decreased and RT increased as the difficulty of the information processing requirements was increased. In addition, performance differences were found between alternative display (e.g., auditory versus visual) and response modalities (e.g., voice, keyboard, microswitch, touch-screen, joystick). For every experimental task, workload ratings tended to follow the same patterns as performance measures: higher levels of subjective workload accompanied poorer performance. In addition, stimulus and response modalities that degraded performance were also rated as having higher workload.

The ratings obtained for the SINGLE-COGNITIVE tasks were either equal to or lower than the overall means (Figure 6a). PE in particular was considered to be very low, reflecting the task characteristics. The ratings were somewhat more variable than the norm, possibly reflecting the diversity of tasks with which they were obtained. Despite this, only three of the rating distributions differed significantly from the "population" distributions: OW, TD and PE. Relatively few scales demonstrated strong statistical relationships with each other. However, TD was highly correlated with ME and FR, and FR was also highly correlated with TP and ST (Table 4). Only TD and ME had correlations that accounted for more than 50 percent of the variance in OW (Table 2b).

SINGLE-MANUAL Category

A variety of one and two-axis tracking tasks were included in this category. As with SINGLE-COGNITIVE, these tasks represented the single-task baseline levels for other categories. The primary source of loading was the physical demands imposed by different experimental manipulations: (1) the bandwidth of the forcing function (three levels in each experiment), (2) order of control (constant or variable), and (3) the number of axes controlled (1 or 2). The display modality was visual, the response modality, manual

Performance and workload levels covaried with the bandwidth manipulations; as bandwidth increased, subjective workload and tracking error increased. In addition, the variable order of control tasks were performed more poorly and were rated as having higher workload. Finally, two-axis tracking was considered to be more loading than one-axis tracking.

In general, SINGLE-MANUAL ratings were higher than the "population" ratings. (Figure 6). FR and ST ratings in particular were higher than for any other tasks, possibly

Table 4: SINGLE-COGNITIVE

Correlations among bipolar ratings

	TD	TP	OP	PE	ME	FR	ST	FA	AT
TP	.47								
OP	.41	.40							
PE	.34	.29	.13						
ME	.74	.49	.40	.36					
FR	.64	.60	.59	.29	.57				
ST	.50	.55	.37	.39	.45	.71			
FA	.34	.43	.28	.35	.28	.52	.54		
AT	.34	.17	.17	.08	.31	.20	.19	.16	
OW	.83	.51	.50	.37	.72	.64	.55	.41	.37

reflecting the subjects' perceptions that some of the conditions were relatively uncontrollable. ME was rated relatively higher than might be expected by the nature of the tasks. AT was rated as "skill-based". The subjects thought their own performance was generally poorer than on other tasks. Most of the rating distributions were significantly different from the "population" distributions except for WWL, ME, PE, and ST. Particularly high correlations among the scales were found between TD and ME, among FR, TP and PE, and among ST, ME, FA and FR (Table 5). As might be expected from the nature of these tasks, a relatively high correlation was found between OW and PE. However, only TD, ME and ST had correlations that accounted for more than 50 percent of the variance (Table 2b).

DUAL-TASK Category

The data from two experiments were included in this category. In each one, continuous one- and two-axis tracking tasks were combined with a discrete, cognitively loading task. Difficulty on the tracking task was manipulated by varying the order of control and bandwidth of the forcing function. For one experiment, the discrete task was three levels of difficulty of an auditory Sternberg memory search task, presented as a pilot's call-sign; responses were vocal. For the other, a spatial transformation task was presented visually or auditorily; responses were vocal or manual. Each task was presented in its single-task form first. The data from these baseline conditions are included in the SINGLE-COGNITIVE and SINGLE-MANUAL categories. The DUAL-TASK conditions represented different combinations of difficulty levels for the two tasks. Time-on-task was manipulated, as well, (ref. II-2) to determine the relationships among fatigue, workload, and event-related cortical potentials in response to the call-signs.

For one experiment, performance on both task components was degraded by time-on-task. Tracking performance was also related to bandwidth. OW, FA, tracking error, and the amplitude of the positive component of the event-related potential were all significantly and positively correlated. For the second experiment (ref. II-15), the visual input modality for the spatial transformation task imposed less workload and interfered less with tracking performance. Speech output resulted in better performance (on both tasks) and less workload than manual output because the latter interfered more with the manual responses required for the tracking task. Subjective ratings were less sensitive to output modality manipulations than to input modality manipulations and to task combinations than individual task levels.

Table 5: SINGLE-MANUAL

Correlations among bipolar ratings

	TD	TP	OP	PE	ME	FR	ST	FA	AT
TP	.49								
OP	.57	.32							
PE	.39	.78	.20						
ME	.75	.39	.44	.29					
FR	.72	.47	.69	.39	.69				
ST	.61	.54	.50	.43	.65	.78			
FA	.39	.34	.35	.32	.42	.54	.67		
AT	.15	.25	.02	.31	.26	.15	.23	.14	
OW	.83	.60	.44	.51	.76	.69	.72	.45	.22

DUAL-TASK ratings were higher, on the average, than the "population" means (Figure 6c). It is not surprising they were higher than the component single task ratings, but it is somewhat surprising that they were higher than the ratings that were given for apparently more complex simulated flying tasks. DUAL-TASK distributions were significantly different from the corresponding "population" distributions for TD, PE, FR, ST, and FA. Among the scales, a few high correlations were notable (Table 6): TD with TP and ME; TP with ME, FR and ST; OP with FR; and FR with ST--patterns almost identical to those observed for the "population". Again, TD, ME and ST were all highly correlated with OW accounting for more than 50 percent of its variance, reflecting a pattern similar to that found for SINGLE-MANUAL. In addition, TP also accounted for more than 50 percent of the variance in OW.

FITTSBERG Category

The FITTSBERG paradigm provides an alternative to the traditional dual-task paradigm in which two unrelated tasks are performed within the same interval. With the FITTSBERG paradigm, the component tasks are functionally related and performed serially: the output or response to one serves to initiate or provide information for the other. A target acquisition task based on FITTS Law (ref. I-9) is combined with a SternBERG memory search task (ref. I-24). Two identical targets are displayed equidistant from a centered probe. Subjects acquire the target on the right, if the probe is a member of the memory set and the target on the left, if it is not. A wide variety of response selection tasks have been used in addition to the Sternberg memory search task: (1) choice reaction time, (2) mental arithmetic, (3) pattern matching, (4) rhyming, (5) time estimation, and (6) prediction. Workload levels for one or both components of the complex task were either held constant or systematically increased or decreased within a block of trials. In addition, the stimulus modality of the two components was the same (visual/visual) or different (auditory/visual).

Response selection performance was evaluated by reaction time (RT) and percent correct. Target acquisition performance was evaluated by movement time (MT). MT but not RT increased as target acquisition difficulty was increased. RT but not MT increased as the cognitive difficulty of response selection was increased. Information sources, processing requirements, and workload levels of the first stage (response selection) appeared to be relatively independent of those for the second stage (response execution), even though some or

Table 6: DUAL-TASKS

Correlations among bipolar ratings

	TD	TP	OP	PE	ME	FR	ST	FA	AT
TP	.72								
OP	.65	.57							
PE	.52	.66	.43						
ME	.83	.70	.59	.46					
FR	.69	.74	.79	.52	.69				
ST	.65	.73	.54	.57	.69	.77			
FA	.33	.42	.50	.40	.34	.59	.49		
AT	.39	.42	.37	.35	.48	.47	.41	.36	
OW	.88	.76	.58	.60	.84	.70	.71	.44	.43

many of the processing stages were performed in parallel, and the activities required for one simultaneously satisfied some of the requirements of the other. Performance decrements were not found for one task component in response to an increase in difficulty of the other. Instead, performance and workload ratings for the combined tasks integrated the component load levels; FITTSBERG ratings and RTs were less than the sum of those for the component tasks performed individually. There was only a small "concurrence" cost of about 40 msec for RT and a 14% increase in ratings for the combined task over single-task baseline levels.

FITTSBERG ratings were generally low except for AT (Figure 6d). The component tasks were not individually difficult and subjects integrated them behaviorally and subjectively, with a consequent "savings" in experienced workload. In addition, rating variability was less than usual. Consequently, all of the rating distributions were significantly different from the "population" distributions.

The following ratings were highly correlated with each other: TD, TP, ME, ST and FR (Table 7). The association between TP and TD is somewhat surprising, as TP is not deliberately manipulated in the FITTSBERG paradigm. The fact that RT was the primary performance metric may have influenced subjects to respond as quickly as possible--a self-imposed time pressure. However, the design of the experimental task did not itself impose time constraints or limits. The low association between OP and OW is also surprising because performance feedback was given frequently. Although TD, TP, ME, and FR were highly correlated with OW, only the correlations between TD and OW, and ME and OW accounted for more than 50 percent of the variance.

POPCORN Category

The POPCORN task is a dynamic, multi-task, supervisory control simulation. It represents operational environments in which decision-makers are responsible for semi-automatic systems. Its name, "POPCORN," reflects the appearance of groups of task elements waiting to be performed (they move around in a confined area and "pop" out when selected for performance). Operators decide which tasks to do and which procedures to follow based on their assessment of the current and projected situation, the urgency of specific tasks, and the reward or penalty for performing or failing to perform them. Simulated control functions provide alternative solutions to different circumstances. They are selected with a magnetic pen and graphics pad and executed by automatic subsystems.

Table 7: FITTSBERG

Correlations among bipolar ratings

	TD	TP	OP	PE	ME	FR	ST	FA	AT
TP	.68								
OP	.38	.39							
PE	.50	.56	.16						
ME	.76	.54	.34	.47					
FR	.69	.67	.45	.44	.63				
ST	.60	.75	.19	.51	.52	.70			
FR	.41	.39	.20	.25	.38	.46	.52		
AT	.36	.17	.05	.23	.42	.20	.15	.13	
OW	.86	.66	.39	.51	.76	.69	.62	.42	.40

Thus, control activities are intermittent and discrete. Task difficulty can be varied by changing the number of tasks, elements/task, scheduled arrival times for successive groups of task elements, speed with which elements move, and penalties imposed for procrastination. The penalties include imposing additional operations or accelerated rates for delayed tasks, deducting points from the score, and losing control over when deferred tasks could be performed.

Experiments conducted with this simulation determined the contributions of different task variables to workload and their behavioral and physiological consequences. Performance was evaluated by examining the score, number of unperformed elements, and completion time. Strategies were evaluated by analyzing the functions selected. Schedule complexity, number of different tasks (rather than the number of elements in each one), and time-pressure-related penalties for procrastination were significantly reflected in the subjective, behavioral, and physiological responses of subjects.

Average rating magnitudes were higher for this group of experiments than for any other (Figure 6e), and their variability was greater. FA was the only factor rated as lower, even though experimental sessions often lasted as long as 5 hours. Distributions of ratings were significantly different from the "population" distributions for every factor except OP. Because TP was the primary way in which workload levels were manipulated, TP ratings were highly correlated with TD, ME, FR, ST, and OW ratings (Table 8) and were considerably higher than the grand mean (46 vs 32).

This task was considered to be the most unpredictable and knowledge-based of the experimental categories (AT = 43 vs 34). PE ratings were higher as well. Even though the computer actually performed the requested functions, virtually continuous selections were required to activate the appropriate functions. This was reflected in a significant correlation between OW and TP. However, PE ratings were not highly correlated with OW across different manipulations. FA and AT were not highly correlated with OW, either, because FA levels were counterbalanced across conditions and AT was relatively constant across all conditions. In this category, only TD and TP accounted for more than 50 percent of the variance in OW.

SIMULATION Category

Three aircraft simulations were combined for this category. Each was conducted in a motion-base general aviation trainer. They were designed to determine the contributions of

Table 8: POPCORN

Correlations among bipolar ratings

	TD	TP	OP	PE	ME	FR	ST	FA	AT
TP	.87								
OP	.68	.69							
PE	.51	.57	.55						
ME	.77	.82	.65	.53					
FR	.65	.66	.74	.51	.58				
ST	.69	.71	.65	.59	.71	.68			
FA	.39	.41	.43	.55	.37	.42	.53		
AT	.27	.25	.16	.22	.30	.26	.24	.14	
OW	.77	.74	.54	.44	.63	.61	.61	.30	.30

individual flight-task components to overall workload and to compare the obtained levels of workload to those predicted by a model. Workload was evaluated by performance on concurrent secondary tasks and ratings. The first experiment (ref. II-8) required control over one (e.g, heading), two (e.g., heading, speed), or three (e.g. heading, altitude, speed) components, with irrelevant dimensions "frozen." As expected, workload increased as the difficulty and complexity of each maneuver increased. The second experiment (ref. II-9) coupled more complex flight-task maneuvers, building up to simulated instrument approaches. Again, workload levels increased as the complexity of flight-task components increased. In the final experiment (ref. II-3), two scenarios, one "easy" and one "hard," were flown. Ratings were obtained during and immediately after each flight. For all three experiments, the various workload measures that were obtained reflected the same underlying phenomena, although the subjective ratings were consistently the most sensitive.

With two exceptions (TP and AT ratings were considerably lower), SIMULATION ratings were similar to the "population" means (Figure 6f). This is surprising, considering the apparently greater magnitude and complexity of task demands imposed on the pilots. In addition, the variability among ratings was the lowest of any category. This might reflect the fact that all of the experimental subjects were instrument-rated pilots familiar with the types of tasks performed. AT was considered to be the most "skill-based" of all of the tasks included in the 16 experiments. Statistical associations among individual scales were lower for this category of experiments than for the rest (Table 9). The highest correlations were found among ME, TD and OP, and among PE, TD, TP, and ST. TD was the only factor that had a strong correlation with OW (accounting for more than 50 percent of its variance).

CONSTRUCTING A WORKLOAD RATING SCALE

Several key points emerged about the subjective experience and evaluation of workload: (1) A phenomenon exists that can be generally termed workload, but its specific causes may differ from one task to the next. (2) Ratings of component factors are more diagnostic than global workload ratings. (3) Subjects' workload definitions differ (thereby contributing to B-S variability); however, the specific sources of loading imposed by a task are more potent determinants of workload experiences than such *a priori* biases. (4) A weighted combination of the magnitudes of factors that contribute to subjects' workload experiences during different tasks provides an integrated measure of overall workload that is relatively stable between raters.

Table 9: SIMULATION

Correlations among bipolar ratings

	TD	TP	OP	PE	ME	FR	ST	FA	AT
TP	.42								
OP	.41	.25							
PE	.46	.61	.25						
ME	.64	.20	.42	.31					
FR	.43	.35	.63	.29	.38				
ST	.53	.64	.38	.60	.36	.58			
FA	.32	.24	.43	.26	.28	.50	.39		
AT	.19	.33	-.13	.24	.02	-.01	.20	-.04	
OW	.86	.36	.38	.42	.65	.33	.45	.21	.08

One of our goals in gathering workload and workload-related ratings, in addition to the information they provided about experimental manipulations, was to amass a data base which would allow us to examine the relationships among different task, behavior, and psychological factors in order to create a valid and sensitive rating technique for subjective workload assessment. Our assumption was that the scale would be multi-dimensional, but that the number of subscales should be less than the number used for research purposes. Thus, the first step was to select the most appropriate set of subscales. The second step was to determine how to combine these subscales to derive a workload score sensitive to different sources and definitions of workload between tasks and raters. The final step was to determine the best procedure for obtaining numeric values for these subscales.

Subscale Selection

We reviewed the information provided by each scale used in the 16 experiments to select the subscales. They should represent the types of phenomena that influence subjective workload experiences in a broad range of tasks (e.g., task-related, subject-related, and performance-related factors), although the importance of individual factors might vary from one type of task to the next. Our goal was to select no more than six factors, so ratings could be obtained during, as well as following, activities performed in operational environments. The following information was considered: (1) sensitivity to differences between tasks (Figure 7), (2) sensitivity to experimental manipulations within tasks(Table 2a), (3) association with subjective ratings of OW (Tables 1b, 3, 4-9), (4) independence from other factors (Tables 1b, 3, 4-9), and (5) subjective importance to raters (Tables 1a, 3; Figure 4a). The following statements about the factors include information drawn from individual experiments, categories of experiments, and the entire data base.

Task-Related Scales

Three of the original scales focused on the objective demands imposed by the experimental tasks. They were TD, TP, and AT.

Task Difficulty. A rating of TD provides the most direct information about subjects' perceptions of the demands imposed on them by a task. TD was considered to be moderately relevant to individual subjects' definitions of workload in the preliminary pairwise comparisons. However, the empirical relationship found between TD and OW ratings was substantially greater than its *a priori* association. In all but one of the 16 experiments, this scale reflected the same experimental manipulations as OW; TD contributed significantly to the OW regression equations in all six categories of experiments. TD was not statistically independent of the other factors that were also found to be important, however. This reduced the information it provided about the workload of different tasks. Although the TD scale was quite sensitive to differences between categories of experiments, its diagnostic value might have been improved if different sources of TD had been distinguished (e.g., mental versus physical).

Time Pressure. TP has been included as a primary factor in most operational definitions and models of workload, where it is quantified by comparing the time required for a series of subtasks to the time available, and it was selected as the factor most closely related to workload in advance of the experiments. However, TP ratings proved to be generally insensitive to manipulations within these experiments. TP ratings were only moderately correlated with OW ratings for individual experiments and categories of experiments. It did discriminate among different types of tasks, however. These findings are due, in part, to the fact that TP was not explicitly manipulated as an experimental variable in many of the experimental tasks. Nevertheless, TP was highly related to more than half of the other variables (the correlation coefficients were greater than 0.70) in 60% of the experiments. It was most closely associated

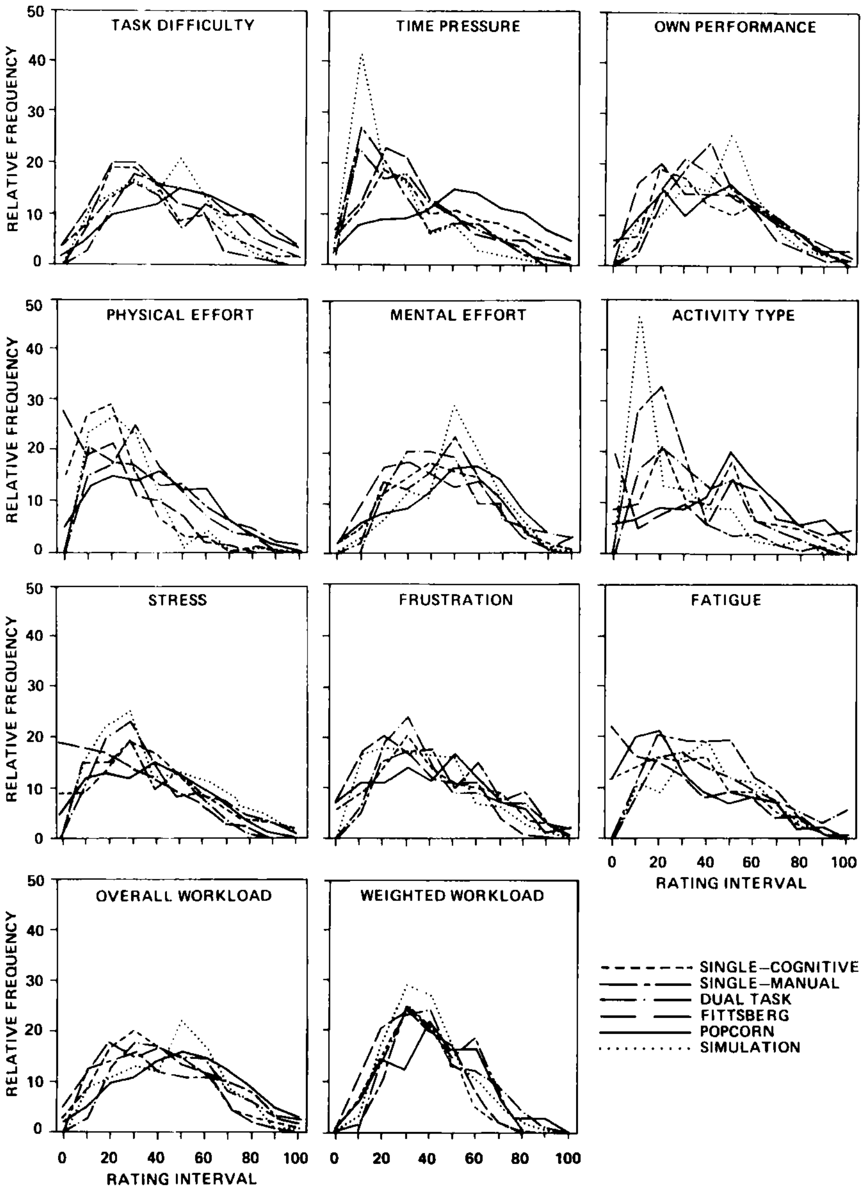


FIGURE 7. RELATIVE FREQUENCY DISTRIBUTIONS OF RATINGS AND WWL SCORES FOR EACH CATEGORY AND SCALE

with PE, ME, FR, and ST--subject-related variables--rather than to the other task-related variables, however. This suggests that perceptions of high or low TP occur because of (and may, in turn, affect) subject-dependent rather than other task-related variables.

Activity Type. Subjects selected AT as a more important contributor to workload than it appeared to be from the empirical results. Furthermore, although AT did discriminate well among categories of tasks, these differences had little or no relationship with their workload levels; the predicted association between skill-based activities and low workload or knowledge-based activities and high workload was not found. AT ratings never correlated significantly with OW and they contributed little to the OW regression equations. Although the *type* of task performed should have some association with the workload it imposes, this scale did not succeed in identifying such a relationship.

Summary of Task-Related Scales. We found that only two task-related scales, TD and TP, provided significant information about workload. Furthermore,we propose dividing the TD scale into two subscales (mental and physical) to identify the specific sources of imposed workload within and between tasks. Thus, three task-related factors were selected: Physical Demands (PD), Mental Demands (MD), and Temporal Demands (TD). These three factors represent the most common ways that workload differences are manipulated across a broad range of activities. They do not represent the cost of achieving task requirements for the operators, however, nor how successful operators were in doing so.

Behavior-Related Scales

The three scales in this category (PE, ME, and OP) provided subjective evaluations of the effort that subjects exerted to satisfy task requirements and opinions about how successful they were in doing so.

Physical Effort. Although PE is a component of most traditional definitions of workload, most of the subjects considered it *a priori* to be essentially unrelated to workload. Empirically, however, this factor discriminated among the different types of experiments and reflected experimental manipulations for tasks with physical demands as a primary workload component. PE ratings were generally low, reflecting the typical nature of laboratory and simulation tasks. Heavy, physical exertion was never required in any of these experiments. PE was not highly correlated with OW within most experiments, however, and did not contribute significantly to the OW regression equation in half of them. It did provide an independent source of information about the subject's experiences, as PE ratings were not highly correlated with ratings of other factors. Its strongest association was with TP (for tasks in which higher levels of imposed TP required higher response rates) and ST (for more complex tasks).

Mental Effort. ME has become an important contributor to the workload of an increasing number of operational tasks because operators' responsibilities are moving away from direct physical control to supervision. *A priori*, ME was considered moderately important to our subjects. Empirically, however, ME ratings were highly correlated with OW ratings in every experimental category and were significantly related to the independent variables in most experiments. ME ratings discriminated among different types of experimental tasks, as well, and it was the second most highly correlated factor with OW. ME ratings were highly correlated with many other task and subject-related variables (e.g., TD, FR, and ST). Thus, the information it provided was somewhat reduced by its lack of independence.

Own Performance. Success or failure in meeting task requirements was considered *a priori* as moderately related to workload. Although OP ratings did not discriminate between types of experimental tasks, it did provide useful and significant information about how the subjects perceived the quality of their performance. OP ratings were significantly correlated with OW ratings in half of the experiments and categories of experiments, and they were relatively independent of other ratings, in comparison to the general finding of high statistical associations.

Summary of Behavior-Related Scales. Although PE and ME each provided significant and relatively independent information about the workload of many experimental tasks, we feel that a single Effort (EF) scale might be sufficient to represent this aspect of workload. This was an arbitrary decision, considering the useful information PE and ME contributed to workload ratings. However, since one of our goals was to reduce the number of bipolar scales, we felt that a combined EF scale could capture the information provided by PE and ME. The additional information in the original PE and ME scales not captured by EF (e.g., the specific source of the load) would be provided by the new MD and PD scales.

Information about the specific source of demands (e.g., physical or mental) can be obtained more directly by asking subjects to evaluate the objective demands that are placed on them than by asking them to introspect about the amount of mental or physical effort exerted. Furthermore, subjective evaluations of task demands can be compared with objective task manipulations for the purpose of validation and prediction. In addition, the B-S variability of ratings for task-related factors should be lower (because the only source of variability would be differences in individuals' sensitivity and understanding), whereas there are at least two interactive sources of variability for behavior-related ratings (the actual levels of effort exerted by each subject, as well as their ability to evaluate these levels introspectively).

The subjects' evaluations of the success or failure of their efforts to accomplish task requirements provided a valuable source of information about workload, because subject's appraisal of performance during a task affects subsequent levels and types of effort exerted. Furthermore, performance decrements observed in operational environments often prompt workload analyses. Thus, some information about performance should be included in any workload assessment technique, even if it is only in the form of a subjective evaluation.

Subject-Related Scales

These scales focused on the psychological impact on the subjects of task demands, behavior, and performance on the subjects. They included FR, ST, and FA.

Frustration. Subjects reported, *a priori*, that FR was the third most relevant factor to workload. Empirically, FR ratings were significantly correlated with OW ratings in most individual experiments and all categories of experiments. FR did not contribute significantly to the OW regression equations, however. This could reflect the fact that FR was not an independent factor; it was strongly correlated with every other factor except AT and PE. FR was only moderately sensitive to experimental manipulations, yet it discriminated among five out of the six categories of experiments. The range of FR ratings across categories was substantial, further suggesting that they provide useful information in distinguishing among types of activities.

Stress. ST has been included in many other subjective rating techniques and is often equated with elevated levels of workload in operational environments. Subjects in these experiments rated ST as the second most important factor in the pretest. Within experiments, ST ratings reflected the same manipulations that influenced OW ratings. However, ST ratings did not discriminate among different types of tasks, it was rarely associated with objective measures of performance and it was the least independent scale (it was highly correlated with every other scale except AT). For this reason, it contributed relatively less to the OW regression equation than its high degree of correlation with OW would suggest.

Fatigue. FA was relatively unrelated to workload in both *a priori* opinions and empirical ratings. Even though the range of FA ratings was the greatest for any scale across categories of experiments (it ranged from 24 to 42), FA ratings rarely covaried with objective performance measures, OW ratings or other factors. One explanation for this lack of relationship could be that fatigue was not manipulated as an experimental variable in most of the studies. In general, it appeared that subjects regarded fatigue as a separate phenomenon from workload.

Summary of Subject-Related Scales. In a multi-dimensional rating technique, it is important to retain some information about the psychological impact on subjects of performing the tasks. Workload, especially the subjective experience of workload, reflects more than the objective demands imposed on an operator. It is apparent from their high intercorrelation, however, that both FR and ST scales are not necessary. ST might be too global a dimension. This term, like workload itself, can mean many different things. The term has been applied to task, environmental, and human phenomena (e.g., heat stress, time stress, emotional stress, physical stress, physiological stress). In fact, an excess of almost any dimension can be termed "stress". FR, in a relatively less ambiguous way, relates task requirements, exerted effort, and success or failure. It provides information about how comfortable operators felt about the effectiveness of their efforts relative to the magnitude of the task demands imposed on them. Although FA can be an experimentally and operationally relevant variable, it was not found to be related to the experience of workload; thus, it was not included as a component of the multi-dimensional rating scale.

Overall Workload Ratings

Although OW ratings were significantly associated with experimental manipulations in most experiments, and distributions of OW ratings were significantly different from one experimental category to the next, the B-S variability within experimental conditions was high; coefficients of variation were often as great as 0.50. In addition, OW ratings appear to reflect different variables in different tasks. Although it is not likely that this contributed to B-S variability within experimental conditions (all subjects experienced the same experimental difficulty manipulations), it does suggest that global workload ratings cannot be compared between tasks. Even though OW ratings provide the most direct and integrated information about the issue in question -- workload -- they may reflect time pressure for one task, variations in effort in another, and different levels of decision making complexity in yet another. Each level of integration has a simplifying effect, reducing complex attributes to progressively more global summaries. There is a point where higher levels of integration cease to provide useful summarization and begin to mask important underlying phenomena. A global workload rating may represent such a point. The component scales can identify variations in sources of loading, as well as their magnitudes, and a weighted combination of them was shown to provide a more stable measure of OW than the global scale itself. This suggests that it is not necessary to obtain a specific OW rating as long as the appropriate components are rated and can be combined.

Weighted Workload Score

The weighted averaging procedure succeeded in reducing B-S variability for all experimental conditions. However, the general information that was obtained in the pretest about differences in workload definition were not sufficient to characterize the specific experiences of subjects that were unique to individual experimental situations. Thus, the WWL score did not achieve the desired level of improvement in statistical sensitivity to experimental variables. Subjective estimates of weighting parameters would have been more useful had they been obtained with reference to a specific experience (e.g., the experimental task) than in the abstract. Self-evaluations obtained in a context are preferable because they provide direct information about the interaction of factors within that context (ref. I-1). and it is this that determines the level of workload.

Verification of Selected Subscales

The high correlations between many of the factors and OW within different categories indicate that multiple dimensions are required to represent the workload of different types of tasks. There is a generic component of workload across tasks as reflected in the correlations of TD, FR, ST, and ME with each experimental category. The task-specific component of workload that is present in some tasks and not in others is reflected in TP and PE. One factor (OP) is moderately related throughout the different types of tasks but is never a primary contributor to workload. The other two factors (FA and AT) are generally unrelated within and between tasks, and consequently were excluded from the new set of subscales.

Before selecting the final set of subscales, several additional analyses were performed. The scales were rank-ordered from most to least relevant: TD, FR, TP, ME, PE, OP, ST, FA, AT. Three scales were eliminated (ST, FA, and AT), and two were combined (EF = ME and PE). The five remaining scales were regressed on OW (Table 10). The percent of variance accounted for by these six scales did not decrease by more than .02 from the variance accounted for by the original nine scales for any of the six categories. The proposed division of TD into Mental (MD) and Physical Demands (PD) could not be simulated with the existing data base.

We examined the three subscales in our data base that are similar to those used in another popular multi-dimensional rating scale, the Subjective Workload Assessment Technique (SWAT) to determine whether these factors alone might provide sufficient information. With the SWAT technique, a preliminary card-sort is performed by each subject to rank-order 27 combinations of three levels (low, medium, high) of the three factors (time load, psychological stress, and mental effort) with respect to the importance they place on them in their personal definition of workload (refs. I-6, I-7, I-21). Conjoint analysis techniques are applied to provide an interval scale of overall workload tailored for individual differences in definition. Subjects provide ratings of low, medium, or high for the three factors following the performance of each experimental task. A single rating of overall workload is obtained by referring to the position on the interval scale identified by that combination of values.

It appears that one of the key assumptions of conjoint analysis (i.e., statistical independence among the components) was not supported by the data from these experiments; ratings of TP, ME, and ST were highly interrelated. Correlations between TP ratings and ST ratings

Table 10

Beta weights for a subset of rating scales regressed on OW (*=p<.01)

	r^2	TD	TP	OP	EF	FR
SINGLE-COGNITIVE	.74	.59*	.06*	.14*	.18*	.04
SINGLE-MANUAL	.79	.54*	.10*	-.12*	.28*	.15*
DUAL-TASKS	.84	.54*	.10*	-.10*	.32*	.11*
FITTSBERG	.78	.60*	.04	.04	.22*	.10*
POPCORN	.64	.52*	.25*	-.15*	.00	.22*
SIMULATION	.75	.77*	.04	.06	.18*	-.10*

were 0.50 or greater, between TP and ME were 0.65 or greater, and between ME and ST were 0.45 or greater in all experiments. For many experiments, correlations were 0.70 or higher. Furthermore, it appears that these three factors alone are not sufficient to represent the range of factors that contribute to workload for a broad range of experimental and operational tasks, as mentioned above.

From a practical, rather than a psychometric, point of view, the independence of workload-related factors presents less of a problem. First, for factors that are both highly related to each other and reflect experimental manipulations, their shared contribution to a weighted estimate of overall workload is simply enhanced, reflecting the actual situation. Second, behavior-related and subject-related factors necessarily reflect task-related factors. Yet task-related factors alone do not provide information about the behavioral and psychological responses of individuals to imposed demands, each important contributors to overall workload. For example, the demand imposed on subjects may be extremely high, yet they may mitigate the levels of workload actually experienced by shedding tasks, lowering their performance standards, or refusing to exert greater and greater levels of effort as task demands increase beyond a certain level. Thus, evaluation of subjects' responses to a task can provide additional information (even though the behavior occurred in response to these demands) as well as highly correlated information. Finally, these scales *can* be driven independently, even though there is often no experimental reason to do so.

Combination of Subscales

Each of the selected subscales provides useful and relevant information about different aspects of subjects' experiences. However, a summary estimate of the overall workload of a task is often needed. Since single OW ratings have been found to be quite variable among subjects and may reflect different factors across tasks, the idea of combining weighted ratings on subscales was suggested as an alternative. However, the weighting procedure adopted for this set of experiments succeeded only in reducing B-S variability. It did not provide estimates of workload that were substantially more sensitive to experimental manipulations than the global OW ratings. Similar sensitivity problems have been found with the SWAT technique. It, too, relies on *a priori*, global judgements about the importance of different factors rather than on the subjective importance of specific variables *within* the target activity to reduce B-S variability. However, B-S variability is often very high for SWAT ratings. Standard deviations that are greater than 50% of the average magnitudes of ratings have been reported in a number of experiments (ref. I-4, II-14, II-15). Despite the relative success of both techniques in identifying variations in workload associated with most experimental manipulations and obtained performance, neither scale has been able to account for a substantial percentage of the variance. For example, a tracking task bandwidth manipulation resulted in highly significant differences in performance, yet accounted for only 8.96% of the WWL score variance and 6.16% of the SWAT ratings (refs. II-14). Even though the former was statistically significant and the latter was not, neither represents the level of sensitivity required for a valid workload assessment technique.

Quantification

Taking into account the results of these and other experiments, it is clear that using the *a priori* biases of subjects about workload to weight or organize subscale ratings into a single workload value may not provide a sufficiently sensitive subjective rating technique. The element missing from both SWAT and the WWL score is information about the sources of workload for the specific task to be evaluated. Regardless of how individuals might personally define workload, workload is caused by different factors from one task to the next and subjects are sensitive to factors that are included in, as well as excluded from, their workload definition. These may take precedence over their natural inclinations to weigh one factor more heavily than another. Since the workload of a task represents the weighted combination of factors that are subjectively relevant during the performance of that task, the weighting function must include information about the sources of loading specific to that task, as well as *a priori* subjective biases. The task-related drivers of subjective experiences should be consistent across individuals who perform the same task. Thus, they should not increase B-S variability within experimental conditions. They do, however, affect the meaning of workload ratings from one task to the next. By enhancing the contribution of factors that are most salient in a particular task to the summary score, its sensitivity should be enhanced.

Figure 8: NASA-TLX RATING SCALE DEFINITIONS

Title	Endpoints	Descriptions
MENTAL DEMAND	*Low/High*	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	*Low/High*	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
TEMPORAL DEMAND	*Low/High*	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
PERFORMANCE	*good/poor*	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
EFFORT	*Low/High*	How hard did you have to work (mentally and physically) to accomplish your level of performance?
FRUSTRATION LEVEL	*Low/High*	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Using the set of six subscales proposed earlier (Figure 8) to represent the possible sources of workload, the following approach might be taken based on the model of the psychological structure of subjective workload estimation presented in Figure 2. For each task (or set of similar tasks), the contribution of each factor to its overall workload could be determined. Although these values could be assigned by an experimenter, the information that is needed relates to the subjective importance of the factors (w), rather than simply their objective contribution (I), as it is the former that influences workload experiences most directly. The simplest way to obtain information about subjective importance would be to ask subjects to assign values to each of the six scales (MD, PD, TP, FR, OP, EF) after a task or set of similar tasks is performed. The same pair-wise comparison technique used in computing the weights for the WWL score could be adopted. Fifteen comparisons would be required to decide which member of each pair of the six factors was most significant in creating the level of workload experienced in performing a particular task. The decision-making process is relatively simple from the subject's perspective and is less tedious than the 36 comparisons used for the 9-factor scale or the 27-factor rank-order used with SWAT. These values would be used to weight the magnitude ratings obtained for the six scales after each experimental condition. The advantage of task-specific weights is that the two sources of variability in ratings that have been identified within tasks (subject's workload definitions) and between tasks (task-related differences in workload drivers) would be represented from the perspective of the raters. The alternatives of using weights provided by the creator of the task to represent the intended sources of loading, or weights that represent nonspecific subject biases, each ignore one potential source of rating variability. A specific example of the proposed rating scale may be found in Appendix B. It summarizes the rating scale descriptions and format, the pairwise technique for determining the subjective importance of each factor in a specific task, and a numerical example of the weighting procedure applied to ratings for two difficulty levels of one task.

Rating scales typically consist of an ordered sequence of response categories that are closed at both ends. End anchors are usually given to provide a frame of reference and to define the correspondence between stimuli (workload experiences) and responses (rated levels). Thus, ratings represent comparative judgements against these extreme values. Our approach has been to ask subjects to provide ratings along a 12-cm line bounded by bipolar adjectives. The anchors are designed to have natural psychological meaning rather than arbitrary values, and to exceed the likely range of rated experiences to avoid the nonlinearities observed for extreme values. Anderson (ref. I-1) and others have suggested that this type of "graphical" format is preferable to discrete categories. The responses were quantified during data analysis by assigning values that ranged from 1 to 100. The resulting values did not represent a ratio scale, and may not have provided even interval data. However, rating variability was acceptably small, most of the scale range was used across tasks, and the numerical values were reliably correlated with experimental manipulations.

The SWAT technique allows only three discrete values to be assigned to each factor--low, medium or high--although reference to a scale provided by the conjoint analysis procedure gives interval workload ratings that range from 1-100. The use of only three scale values is understandable from a practical point of view (a greater number would make the initial sorting procedure nearly impossible), however, it significantly reduces the sensitivity of this technique. The workload of most tasks lies somewhere in the mid-range, and subjects often avoid giving extreme values. Furthermore, scales with fewer than six or seven increments are particularly susceptible to response nonlinearities near the endpoints and, in addition, there are distribution effects (ref. I-1). Furthermore, SWAT uses word labels for each interval, which may be risky because each may connote unequal subjective category widths (ref. I-1). The strength of the SWAT technique lies in the fact that it provides an interval scale of workload by virtue of the conjoint analysis technique employed. Although the benefits of this are clear from a psychometric point of view, the practical cost of the procedure and the limitations it imposes on the range of rating values limits its utility. This is particularly true given the high B-S variability observed in the ratings.

Thus, our recommendation is that a fairly wide range of increments is desirable. Anderson (ref. I-1) suggested than the optimal range of rating steps is from 10 to 20. With more steps, ratings tend to cluster because subjects provide ratings in round numbers and are not sensitive to very fine distinctions. Furthermore, graphic ratings that are quantified on a scale from 1-100 with 1-point increments suggest greater sensitivity to experimental manipulations than subjects are likely to be capable of producing. Discrete numeric ratings could be obtained verbally (e.g., 0-20) during an operational task where it is not practically possible to present an analog scale for rating each factor on a computer display or paper-and-pencil form. However, graphic scales, represented by an unmarked continuum bounded by extreme anchor values, are preferable. This continuum can be divided into equal intervals during data analysis for scoring.

Reference Tasks

A final point will be considered briefly: the additional reduction in B-S variability that can be obtained with the introduction of a reference task. It is unlikely that workload ratings are given absolutely or in reference to a global internal scale of workload that can be applied equally to all tasks. Rather, subjects compare the current situation with similar experiences and evaluate its workload with reference to the ranges and magnitudes of common features; each subject may select different reference activities unless one is explicitly provided. Furthermore, experimental conditions are often presented in a counter-balanced order, and the progression of task difficulties from easy to hard or vice versa may influence the subjective anchor points used in providing ratings differently. This source of rating variability is not obvious from the ratings that are provided. Thus, even without an explicit reference task, presenting experimental subjects with illustrative examples of the range and average difficulties of the tasks to be evaluated helps provide a stable judgemental set and orients the subject to the types of tasks to be performed (ref. I-1).

The use of reference tasks for workload ratings was suggested by Gopher (refs. I-10, I-11). His initial suggestion was that a single task could be presented as a common reference within and between experiments. It could be assigned an arbitrary value and the workload levels of the remaining tasks rated with respect to this task. The initial hope was that one task could be used as a reference for a wide range of different tasks. The goal was to discover an underlying psychophysical function analogous to that existing for many perceptual processes involving objective, physical stimuli. He found, as we did, that the workload of different tasks may be caused by different factors. Thus, reference tasks must be selected that share elements in common with the experimental tasks. When this is done, ratings can be assigned to similar tasks in comparison with a common activity. This approach could be coupled with the rating technique suggested above. The reference task could be used to obtain subjective estimates of the importance of the six workload-related factors for that type of activity. These weights could be applied to each member of a set of experimental tasks in which the magnitudes of different factors were experimentally varied. This would have the practical advantage of reducing the number of times importance weights would have to be obtained, and it would emphasize the salient characteristics of the reference task. The disadvantage of obtaining factor weights for groups of tasks is the possibility that the subjective importance of the factors might interact with variations in their magnitudes from one task to the next. This procedure would still be preferable to unweighted ratings or *a priori* weights based on abstract features or levels.

The great success of the Cooper-Harper Rating Scale for Aircraft Handling Qualities (refs. I-3, I-29) suggests the additional value of providing concrete examples of scale values. Test pilots use this rating procedure to provide subjective evaluations of the handling qualities of aircraft and aircraft simulations. They are "calibrated" by experiencing different levels of aircraft handling qualities in variable stability aircraft. This provides concrete experiences as references for each of the 10 scale values. By providing examples of tasks designated as low or high workload, B-S rating variability could be reduced.

Validation

An extensive validation study was completed recently to determine (1) whether the six NASA-TLX subscales are adequate to characterize variations in the sources of workload among different tasks, (2) whether the weights obtained from subjects are diagnostic with respect to the source of workload unique to each task, and (3) whether the task-related weighting procedure provides a global workload score that is sensitive to workload variations within and between tasks. Thirteen different experimental tasks were presented to a group of six male subjects. Blocks of experimental trials were repeated at least eight times per task, although many were repeated more often to present different experimental manipulations

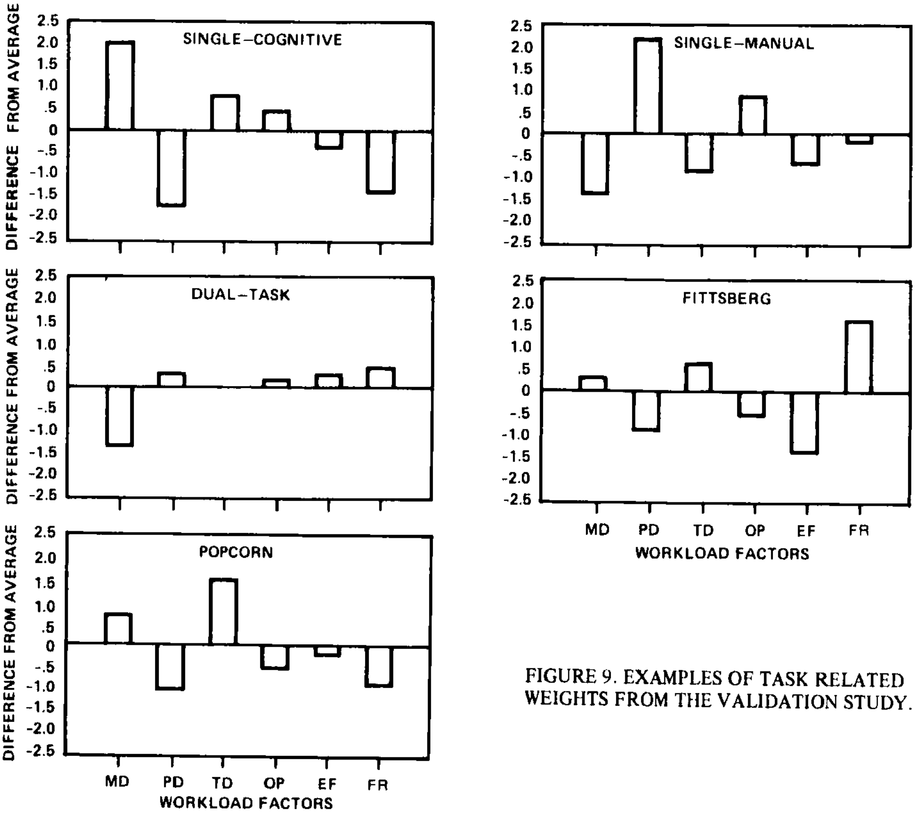


FIGURE 9. EXAMPLES OF TASK RELATED WEIGHTS FROM THE VALIDATION STUDY.

within a task. The tasks included manual control (one axis compensatory tracking, subcritical instability tracking, step tracking, target acquisition), perception (iconic memory, pattern recognition), short-term memory (the Sternberg task, serial pattern matching), cognitive processing (mental rotation, logical reasoning, serial arithmetic, time production), parallel and serial dual-tasks (variations of FITTSBERG, two axis compensatory tracking), and the POPCORN supervisory control task. The experimental tasks were grouped according to the categories in the initial data base: (1) SINGLE-COGNITIVE, (2) SINGLE-MANUAL, (3) DUAL-TASK, (4) FITTSBERG, (5) and POPCORN. The SIMULATION category was not included. The initial results will be discussed very briefly to illustrate the success of the proposed rating scale in meeting its objectives. A more complete description of the experimental tasks, procedure, and results is in progress.

Weights

Subjects were able to specify which factors contributed most (and least) to the workload they experienced during each type of task. As an example the weights given for one task selected from each category are depicted in Figure 9. The workload sources for one of the tasks in each category (weights) are represented as deviations from an "average" weight of 2.5. The values each weight could attain ranged from 0 to 5 (not at all important to more important than any other factor, respectively). The subjective evaluations of the contribution of different sources of workload varied significantly among the different types of tasks. These evaluations reflected the objective experimental manipulations (e.g., MD, PD, and TD) as well as the subjects' individual responses to them (e.g., OP, EF, FR). For example, MD was the most significant contributor to the workload of the logical reasoning task, while PD was the most significant contributor to the workload of the subcritical instability tracking. For different tasks that shared common sources of loading, similar patterns of weights were found. For example, MD was the primary source of workload for SINGLE-COGNITIVE tasks that

Table 11: Validation Study

Correlations among bipolar ratings

	MD	PD	TD	OP	EF	FR
PD	.57					
TD	.58	.50				
OP	.36	.27	.32			
EF	.76	.58	.66	.40		
FR	.54	.44	.52	.57	.69	
OW	.84	.70	.67	.46	.84	.70

Table 12: Validation Study

Beta weights for the six rating subscales regressed on OW (*=p<.01)

	r^2	MD	PD	TD	OP	EF	FR
SINGLE-COGNITIVE	.88	.43*	.15*	.04	.01	.33*	.13*
SINGLE-MANUAL	.78*	.38*	.39*	.11*	.12*	.21*	.00
DUAL-TASKS	.82	.41*	.19*	.02	.09*	.29*	.20*
FITTSBERG	.86	.32*	.24*	.17*	.09*	.16*	.19*
POPCORN	.90	.34*	.23*	.22*	.03	.19*	.10*
OVERALL	.86	.38*	.22*	.08	.05	.24*	.16*

had no time constraints, whereas both MD and TD were equally important for SINGLE-COGNITIVE tasks that placed time limits on information gathering, processing, or response.

When weights were obtained several times for the same task, the relative importance of task-related factors did not change significantly, although the importance of the subjects' emotional responses to the task (e.g., FR) was reduced as task performance improved through training. When weights were obtained for different components of a complex task, they distinguished among the sources of load unique to each task component as well as for the combined tasks.

It is clear from the results of analyses performed on the weights, that the sources of load do, indeed, vary among tasks (at least from the perspectives of the raters). Although these weights still reflect some individual differences in the subjective importance of different factors, the variations in sources of workload characteristic of different types of activities provides a more potent description of the task characteristics than could the *a priori* weights obtained from each rater. It is likely that these differences should be taken into account when computing a weighted average. Furthermore, the values assigned to each factor averaged across subjects provided a diagnostic tool. By identifying the specific source of workload in a task it provides a basis for deciding how to modify unacceptably high levels of workload in operational environments.

Ratings

As we found with the initial set of nine scales, ratings on some of the six NASA-TLX subscales were significantly correlated (Table 11); however, the six subscales appeared to be somewhat more independent than were the original nine scales. For some factors (e.g., TD and FR) magnitude ratings were highly correlated with the subjective importance placed on that factor as a source of workload. For example, time pressure was a significant source of workload only when it was high. When MD or PD was a primary source of workload, however, the **magnitude** ratings were not necessarily high. For example, PD was considered to be the primary source of load for the subcritical tracking task, yet PD ratings were quite low (26). Many tasks were thought to have MD as a primary source of workload, yet MD ratings ranged from 20 to 66, depending on the magnitude of the mental demands each task placed on the subjects. EF was considered to be a moderately important source of workload (weights varied from 1.2 to 2.8) for every task and EF ratings were consistently highly correlated with OW ratings. The importance of OP varied widely across tasks (weights varied from .8 to 3.3), yet OP ratings were relatively unrelated to OW ratings. As expected, the sensitivity of individual scales to experimental manipulations varied depending on the sources of load and ranges of levels in each task.

As with the initial data base, ratings on the six NASA-TLX subscales were regressed against OW ratings within each category and across categories. Table 12 shows that these six scales were able to account for a highly significant percentage of the variance in OW ratings (r-squared values ranged from 0.78 to 0.90), even though their numbers was reduced from the original nine. In addition, the correlation among the regression coefficients were rarely significant, providing additional evidence that these six scales represent relatively independent sources of information about the workload imposed by different tasks.

Within each experiment, the B-S variability in the magnitude of the WWL ratings for the six subscales was generally less than the B-S variability of global OW ratings. In contrast to the subject-related weights used in the previous set of experiments, however, the task-related weights provided workload estimates that were more sensitive to experimental manipulations than the global workload OW ratings were. When TD, MD or PD was varied within a task the ratings obtained for these factors were significantly different. Since these factors were also weighted more heavily in computing the averaged weighted workload score, the sensitivity of the summary value was enhanced as well. Highly significant differences in subjective workload ratings were found within each experiment that reflected meaningful experimental manipulations which covaried with objective performance measures. Using the POPCORN tasks as an example, both the rate of movement of task elements and the inter-arrival rate of groups of elements resulted in highly significant differences among scores. Average scores ranged from 200 to 700 between the most difficult and the easiest versions while average workload ratings ranged from 47 to 73 for the same experimental conditions. On the other hand, where performance differences were not found (e.g.. among replications once asymptotic performance levels were reached), subjective workload measures were not significantly different.

In a different study, we looked at the effect of administering the NASA-TLX either verbally, by paper-and-pencil, or by computer. Subjects provided TLX ratings following asymptotic performance of two levels (E,H) of three tasks (target acquisition, grammatical reasoning, and unstable tracking) using the three methods. On the average, ratings obtained by the computer method were 2 points higher than by the verbal method, and 7 points higher than by the paper-and-pencil method. Although the ratings obtained by the computer method were significantly different than those obtained by the the paper-and-pencil method, the absolute differences in numbers are less important than the fact that the patterns in the magnitudes of the ratings were extremely consistent for all tasks. The correlations among the three methods were very high: computer vs verbal = .96, computer vs paper/pencil = .94, and verbal vs paper/pencil =.95.

This study was conducted again four weeks later to evaluate the test/retest reliability in the rating techniques. The relationships among the three methods were the same in the initial test as in the retest: there were no significant differences between ratings given for a task in the initial test and ratings for that same task in the retest, for any of the three methods. The correlation between the test/retest ratings was .83. Despite the consistency in the patterns of ratings in the three methods, we feel the verbal method is the least desirable method, even though it is the easiest to administer. In particular, confusion can arise due to population stereotypes about whether ones own performance should have a high number associated with good performance and a low number associated with bad performance. In the TLX scale, good performance is associated with a low number, as lower workload is usually accompanied by better performance.

SUMMARY

This chapter has presented the rationale behind the design of the NASA-TLX for subjective workload assessment based on the results of a three-year research effort. Given the many problems outlined above, the ability of subjects to give meaningful ratings is remarkable. Because this area has received relatively little theoretical attention, our goal was to provide a data base containing examples of a wide variety of activities from which general principles and relationships could be drawn.

Until recently, subjective ratings have been treated as tools that are subject to undesirable biases and that represent the discredited practice of Introspection. Instead, it appears that the biases observed in workload ratings, as for subjective evaluations of other factors, may actually reflect interesting and significant cognitive processes (ref. I-1). At least five sources of rating variability were identified: (1) variations in the objective and subjective importance of different features to the workload of different tasks; (2) experimental variations in the magnitudes of different factors; (3) differences in the rules by which individuals combine information about the task, their own behavior, and psychological responses to the task into subjective workload experiences; (4) difficulties associated with translating a subjective experience into an overt evaluation; and (5) lack of sensitivity to experimental manipulations or psychological processes. To some extent, these variables are under experimental control. However, the subjective experience of workload represents the intersection between objective task demands and each individual's response to them. Thus, uncontrolled sources of variability are necessarily introduced. Differences in workload associated with the specific composition of a task and its psychological counterpart can be identified though subjective reports about specific (rather than abstract or general) activities. This information is included in the proposed multi-dimensional rating scale, NASA-TLX, in the form of weights applied to ratings for specific factors. The last two sources of variability, those related to psychometric and sensitivity problems, are likely to remain as uncontrolled and undesirable sources of rating variability. However, by soliciting appropriate subscales, weighting factors, scale designs, and reference tasks, there should be a sufficient improvement in sensitivity and stability so that these other sources of variability should only add "noise" rather than compromise the utility of subjective ratings as a significant and practical source of information about workload.

From all of the information obtained in the initial analysis of the original data base and from the preliminary analysis of the set of experiments included in the validation study, it appears that the NASA-TLX scale is more sensitive to experimental manipulations of workload than either a global rating or a combination of subscales weighted to reflect the *a priori* biases of the subjects only. Furthermore, each of the six subscales was found to be the primary source of loading in at least one experiment and to contribute to the workload of others. Each factor was, therefore, able to contribute independent information about the structure of different tasks. Thus, NASA-TLX provides additional information about the tasks that is not available from either SWAT or the original, nine-factor scale.

NASA-TLX ratings were obtained quickly (it took less than one minute to obtain the six ratings after each experimental condition). In addition, it took no more than two minutes to obtain the weights for each different type of task. This suggests that the proposed multi-dimensional rating scale would be a practical tool to apply in operational environments (which the nine-factor scale was not) and data analysis is substantially easier to accomplish than it is with SWAT, which requires a specialized conjoint analysis program. The weighted combination of factors provides a sensitive indicator of the overall workload between different tasks and among different levels of each task, while the weights and the magnitude of the ratings of the individual scales provide important diagnostic information about the specific source of loading within the task.

APPENDIX A: Sample Application of the NASA-TLX.

EXAMPLE:

COMPARE WORKLOAD OF TWO TASKS THAT REQUIRE A SERIES OF DISCRETE RESPONSES. THE PRIMARY DIFFICULTY MANIPULATION IS THE INTER-STIMULUS INTERVAL (ISI) – (TASK 1 = 500 msec. TASK 2 = 300 msec)

PAIR-WISE COMPARISONS OF FACTORS:

INSTRUCTIONS: SELECT THE MEMBER OF EACH PAIR THAT PROVIDED THE MOST SIGNIFICANT SOURCE OF WORKLOAD VARIATION IN THESE TASKS

PD / [✓ MD]
[✓ TD] / MD
OP / [✓ MD]
FR / [✓ MD]
[✓ EF] / MD

[✓ TD] / PD
[✓ OP] / PD
[✓ FR] / PD
[✓ EF] / PD
[✓ TD] / OP

[✓ TD] / FR
[✓ TD] / EF
OP / [✓ FR]
OP / [✓ EF]
EF / [✓ FR]

TALLY OF IMPORTANCE SELECTIONS

MD III	=	3
PD	=	0
TD IIIII	=	5
OP I	=	1
FR III	=	3
EF III	=	3
SUM	=	15

RATING SCALES:

INSTRUCTIONS: PLACE A MARK ON EACH SCALE THAT REPRESENTS THE MAGNITUDE OF EACH FACTOR IN THE TASK YOU JUST PERFORMED

DEMANDS	RATINGS FOR TASK 1:			RATING		WEIGHT		PRODUCT
MD	LOW	I ‗‗x‗‗‗‗‗‗‗ I	HIGH	30	X	3	=	90
PD	LOW	I ‗x‗‗‗‗‗‗‗‗ I	HIGH	15	X	0	=	0
TD	LOW	I ‗‗‗‗‗x‗‗‗‗ I	HIGH	60	X	5	=	150
OP	EXCL	I ‗‗‗x‗‗‗‗‗‗ I	POOR	40	X	1	=	40
FR	LOW	I ‗‗x‗‗‗‗‗‗‗ I	HIGH	30	X	3	=	90
EF	LOW	I ‗‗‗x‗‗‗‗‗‗ I	HIGH	40	X	3	=	120
				SUM			=	490
				WEIGHTS (TOTAL)			=	15
				MEAN WWL SCORE			=	$\boxed{32}$

DEMANDS	RATINGS FOR TASK 2:			RATING		WEIGHT		PRODUCT
MD	LOW	I ‗‗x‗‗‗‗‗‗‗ I	HIGH	30	X	3	=	90
PD	LOW	I ‗x‗‗‗‗‗‗‗‗ I	HIGH	25	X	0	=	0
TD	LOW	I ‗‗‗‗‗‗x‗‗‗ I	HIGH	70	X	5	=	350
OP	EXCL	I ‗‗‗‗x‗‗‗‗‗ I	POOR	50	X	1	=	50
FR	LOW	I ‗‗‗‗x‗‗‗‗‗ I	HIGH	50	X	3	=	150
EF	LOW	I ‗x‗‗‗‗‗‗‗‗ I	HIGH	30	X	3	=	90
				SUM			=	730
				WEIGHTS (TOTAL)			=	15
				MEAN WWL SCORE			=	$\boxed{49}$

RESULTS:

SUBSCALES PINPOINT SPECIFIC SOURCE OF WORKLOAD VARIATION BETWEEN TASKS (TD). THE WWL SCORE REFLECTS THE IMPORTANCE OF THIS AND OTHER FACTORS AS WORKLOAD-DRIVERS AND THEIR SUBJECTIVE MAGNITUDE IN EACH TASK

REFERENCES I

[1] Anderson, N. H. (1982). *Methods of Information Integration Theory.* New York: Academic Press.

[2] Childress, M. E., Hart, S. G. & Bortolussi, M. R. (1982). The reliability and validity of flight task workload ratings. *Proceedings of the Human Factors Society 26th Annual Meeting.* Santa Monica, CA: Human Factors Society, 319-323.

[3] Cooper, G. E. & Harper, R. P. (1969). The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities (NASA TN-D-5153) Washington, D.C.: National Aeronautics and Space Administration.

[4] Courtright, J. F. & Kuperman, G. (1984). Use of SWAT in USAF System T & E. *Proceedings of the Human Factors Society 28th Annual Meeting.* Santa Monica, CA: Human Factors Society, 700-704.

[5] Damos, D. L. (1984). Classification systems for individual differences in multiple-task performance and subjective estimates of workload. *Proceedings of the 20th Annual Conference on Manual Control.* (NASA-CP 2341) Washington, D.C.: National Aeronautics and Space Administration, 97-104.

[6] Eggemeier, F. T. (1981). Current issues in subjective assessment of workload. *Proceedings of the Human Factors Society 25th Annual Meeting.* Santa Monica, CA: Human Factors Society, 513-517.

[7] Eggemeir, F. T., Crabtree, M. S.. Zingg, J. J., Reid, G. B., & Shingledecker, C. A. (1982). Subjective workload assessment in a memory update task. *Proceedings of the Human Factors Society 26th Annual Meeting.* Santa Monica, CA: Human Factors Society. 643-647.

[8] Ericsson, K. A. & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, **87** (3), 215-251.

[9] Fitts, P. M. & Peterson, J. R. (1964). Information capacity of discrete motor responses. *Journal of Experimental Psychology,* **67** 103-112.

[10] Gopher, D., & Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? *Human Factors*, **26** (5), 519-532.

[11] Gopher, D., Chillag, N. & Arzi, N. (1985). The psychophysics of workload - A second look at the relationship between subjective measures and performance. *Proceedings of the Human Factors Society 29th Annual Meeting.* Santa Monica, CA: Human Factors Society, 640-644.

[12] Hart, S. G. (1986). Theory and measurement of human workload. In J. Zeidner (Ed.) *Human Productivity Enhancement*. New York : Praeger, 496-555.

[13] Hart, S. G., Childress, M. E., & Hauser, J. R. (1982). Individual definitions of the term "workload". *Proceedings of the 1982 Psychology in the DOD Symposium*. USAFA, CO, 478-485.

[14] Hauser, J. R., Childress, M. E. & Hart, S. G. (1983). Rating consistency and component salience in subjective workload estimation. *Proceedings of the 18th Annual Conference on Manual Control*. (AFWAL-TR-83-3021) Wright-Patterson Air Force Base, OH, 127-149.

[15] Johanssen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A. & Wickens, C. (1979). Final report of experimental psychology group. In N. Moray (Ed.), *Mental Workload: Its Theory and Measurement*. New York: Plenum Press, 101-116.

[16] Madni, A. & Lyman, J. (1983). Model-based estimation and prediction of task-imposed mental workload. *Proceedings of the Human Factors Society 27th Annual Meeting*. Santa Monica, CA: Human Factors Society, 314-318.

[17] Mane, A, M. (1985). Adaptive and part-whole training in the acquisition of a complex perceptual-motor skill. Unpublished thesis.

[18] Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, **84** (3), 231-259.

[19] Rasmussen, J. (1983). Skills, rules, and knowledge; Signals, signs, and symbols, and other distinctions in human performance models. *IEEE Systems, Man, and Cybernetics*, New York: Institute of Electrical and Electronic Engineers, 257-266.

[20] Reid, G. B., Shingledecker, C. A, Nygren, T. E. & Eggemeier, F. T. (1981). Development of multidimensional subjective measures of workload. *Proceedings of the International Conference on Cybernetics and Society*, New York: Institute of Electrical and Electronic Engineers, 403-406.

[21] Reid, G. B., Eggemeier, F. T., & Nygren, T. E. (1982). An individual differences approach to SWAT scale development. *Proceedings of the Human Factors Society 26th Annual Meeting*. Santa Monica, CA: Human Factors Society, 639-642.

[22] Sheridan, T. B., and Stassen, H. (1979). Toward the definition and measurement of the mental workload of transport pilots. (Final report DOT-OS-70055) Cambridge, MA: MIT.

[23] Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

[24] Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, **30**, *276-315*.

[25] Turksen, I. B. & Moray, N. & Fuller, K. (in press). A linguistic rule-based expert system for mental workload. In H. J. Bullinger & H. J. Warnecke (Eds.) *Toward the Factory of the Future.*

[26] Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124-1131.

[27] US Army Combat Developments Experimentation Center. (1984). *Scout-Observer Unit Test II: Scout II Test Report. (CDEC-TR-84-015)*

[28] White, S. A., McKinnon, D. P., & Lyman, J. (in press). Modified petri net sensitivity to workload manipulations. *Proceedings of the 21st Annual Conference on Manual Control.* Columbus, OH: Ohio State University.

[29] Wierwille, W. W. (1984). *Comparitive Evaluation of Workload Estimation Techniques in Piloting tasks.* (NASA CR-166496) Washington D.C. : National Aeronautics and Space Administration.

[30] Wierwille. W. W., Skipper, J. H. & Rieger, C. A. (1984). Decision tree rating scales for workload estimation: Theme and variations. *Proceedings of the 20th Annual Conference on Manual Control.* (NASA CP 2341) Washington, D.C.: National Aeronautics and Space Administration, 73-84.

REFERENCES II

[1] Battiste, V. & Hart, S. G. (1985). Predicted versus experienced workload and performance on a supervisory control task. *(Proceedings of the Third Biannual Symposium on Aviation Psychology).* Columbus, OH: Ohio State University, 265-262.

[2] Biferno, M. A. (1985). *Mental Workload Measurement: Event-Related Potentials and Ratings of Workload and Fatigue.* (NASA CR-177354) Washington, D. C.: National Aeronautics and Space Administration.

[3] Bortolussi, M. R.. Kantowitz, B. H. & Hart. S. G. (1985). Measuring pilot workload in a motion base trainer: A comparison of four techniques. *Proceedings of the Third Biannual Symposium on Aviation Psychology.* Columbus, OH: Ohio State University, 263-270.

[4] Hart, S. G., Battiste, V., Chesney, M. A., Ward, M. M. & McElroy, M. (in press). Type A vs Type B: comparison of workload, performance and cardio vascular measures

[5] Hart, S. G., Battiste, V. & Lester, P. T. (1984). POPCORN: A supervisory control simulation for workload and performance research. *Proceedings of the 20th Annual Conference on Manual Control.* (NASA CP-2341) Washington, D. C.: National Aeronautics and Space Administration, 431-454.

[6] Hart, S. G., Sellers, J. J. & Guthart, G. (1984). The impact of response selection and response execution difficulty on the subjective experience of workload. *Proceedings of the 28th Annual Meeting of the Human Factors Society.* Santa Monica, CA: Human Factors

Society, 732-736.

[7] Hart, S. G., Shively, R. J., Vidulich, M. A. & Miller, R. C. (in press). The effects of stimulus modality and task integrality: Predicting dual-task performance and workload from single task levels. *Proceedings of the 21st Annual Conference on Manual Control.* Columbus OH: Ohio State University.

[8] Kantowitz, B. H., Hart, S. G., Bortolussi, M. R., Shively, R. J., & Kantowitz, S. C. (1984) Measuring pilot workload in a moving-base simulator: II. Building levels of load. *Proceedings of the 20th Annual Conference on Manual Control.* (NASA CP-2341) Washington, D. C.: National Aeronautics and Space Administration, 359-372.

[9] Kantowitz, B. H., Hart, S. G., Bortolussi, M. R., Shively, R. J. & Kantowitz, S. C. Measuring pilot workload in a moving-base simulator: Building levels of workload. (Unpublished manuscript).

[10] Miller, R. C. & Hart, S. G. (1984). Assessing the subjective workload of directional orientation tasks. *Proceedings of the 20th Annual Conference on Manual Control.* (NASA CP-2341) Washington, D. C.: National Aeronautics and Space Administration, 85-96.

[11] Mosier, K. L. & Hart, S. G. (in press). Levels of information processing in a Fitts Law task. *Proceedings of the 21st Annual Conference on Manual Control.* Columbus OH: Ohio State University.

[12] Shively, R. J. (1985). *Evaluation of Data Entry Devices.* Unpublished masters thesis. West Lafayette, IN: Purdue University.

[13] Staveland, L., Hart, S. G. & Yeh, Y.-Y. (in press). Memory and subjective workload assessment. *Proceedings of the 21st Annual Conference on Manual Control.* Columbus OH: Ohio State University.

[14] Vidulich, M. A. & Tsang, P. S. (in press). Techniques of subjective workload assessment: A comparison of two methodologies. *Proceedings of the Third Biannual Symposium on Aviation Psychology.* Columbus, OH: Ohio State University 239-246. To appear in a special issue of *Ergonomics.*

[15] Vidulich, M. A. & Tsang, P. S. (1985). Assessing subjective workload assessment: A comparison of SWAT and the NASA-Bipolar methods. *Proceedings of the of the Human Factors Society 29th Annual Meeting.* Santa Monica, CA: Human Factors Society, 71-75.

[16] Yeh, Y.-Y., Wickens, C. D. & Hart, S. G. (1985). The effect of varying task difficulty on subjective workload. *Proceedings of the of the Human Factors Society 29th Annual Meeting.* Santa Monica, CA: Human Factors Society, 765-770.

REFERENCES III

Recent Experimental uses of the NASA-TLX

[1] Battiste, V. (1987). *Part-Task vs Whole-Task Training: Twenty years later.* Unpublished Master's Thesis. San Jose State University.

[2] Fuld, R., Liu, Y., & Wickens, C. D. (1987). *Computer monitoring vs self monitoring: The impact of automation on error detection. (ARL-87-3/NASA-87-4). Champaign: University of Illinois, Department of Aviation.*

[3] Johnson, W. W. & Hart, S. G. (in press). Step tracking shrinking targets. In *Proceedings of the 31st Annual Meeting of the Human Factors Society.* Santa Monica: Human Factors Society.

[4] Liu, Y. & Wickens, C. D. (1987). *Mental workload and cognitive task automation: An evaluation of subjective and time estimation techniques.* (ERL-87-2/NASA-87-2). Champaign: University of Illinois, Engineering-Psychology Research Laboratory.

[5] Liu, Y. Wickens, C. D. (in press). The effect of representational code, response modality, and automation on dual-task performance and subjective workload: An integrated approach. In *Proceedings of the 31st Annual Meeting of the Human Factors Society.* Santa Monica: Human Factors Society.

[6] NASA Task Load Index (TLX): *Computerized Version.* (1986). Moffett Field, CA: NASA-Ames Research Center, Aerospace Human Factors Research Division.

[7] NASA Task Load Index (TLX): *Paper-and-Pencil Version.* (1986). Moffett Field, CA: NASA-Ames Research Center, Aerospace Human Factors Research Division.

[8] Nataupsky, M. & Abbott, T. (in press). Comparisons of workload measures on a computer-generated primary flight display. In *Proceedings of the 31st Annual Meeting of the Human Factors Society.* Santa Monica: Human Factors Society.

[9] Pepitone, D. & Shively, R. J. (in press). Predicting pilot workload. In *Proceedings of the 1987 Aerospace Technology Conference and Exposition.* Long Beach, CA: Society of Automotive Engineers.

[10] Shively, R. J., Battiste, V., Hamerman-Matsumoto, J., Pepitone, D. D., Bortolussi, M. R., & Hart, S. G. (in press). Inflight evaluation of pilot workload measures for rotorcraft. In *Proceedings of the Fourth Symposium on Aviation Psychology.* R. Jensen (Ed.). Columbus: Ohio State University.

[11] Tsang, P. S. & Johnson, W. W. (in press). Automation: Changes in cognitive demands and mental workload. In *Proceedings of the Fourth Symposium on Aviation Psychology.* R. Jensen (Ed.). Columbus: Ohio State University.

12 Tsang, P. S. & Vidulich, M. A. (in press). Time-sharing visual and auditory tracking tasks. In *Proceedings of the 31st Annual Meeting of the Human Factors Society.* Santa Monica: Human Factors Society.

13 Vidulich, M. A. & Pandit, P. (in press). Individual differences and subjective workload assessment. In *Proceedings of the Fourth Symposium on Aviation Psychology.* R. Jensen (Ed.). Columbus: Ohio State University.

14 Vidulich, M. A. Pandit, P. (in press). Consistent Mapping and spatial consistency in target detection and response execution. *Proceedings of the Fourth Mid-Central Ergonomics/Human Factors Conference.* Champaign: University of Illinois.

[15] Vidulich, M. A., & Tsang, P. S. (in press). Rating scale and paired comparison approaches to subjective mental workload assessment. In *Proceedings of the 31st Annual Meeting of the Human Factors Society.* Santa Monica: Human Factors Society.

[16] Wild, H. M., Stokes, J., Weiland. W. & Harrington, N. (in press). Experimental evaluation of the submarine localization module for the naval air anti-submarine warfare P3 tactical coordination officer (Technical Report). Warminster, PA: Naval Air Development Center.

statology.org

How to Perform the Wilcoxon Signed Rank Test - Statology

Zach

~4 minutes

The **Wilcoxon Signed Rank Test** is the non-parametric version of the paired t-test. It is used to test whether or not there is a significant difference between two population means.

Use the Wilcoxon Signed Rank test when you would like to use the paired t-test but the distribution of the differences between the pairs is severely non-normally distributed.

The easiest way to determine if the differences are non-normally distributed is to create a histogram of the differences and see if they follow a somewhat normal, “bell-shaped” distribution.

Keep in mind that the paired t-test is fairly robust to departures from normality, so the deviation from a normal distribution needs to be pretty severe to justify the use of the Wilcoxon Signed Rank test.

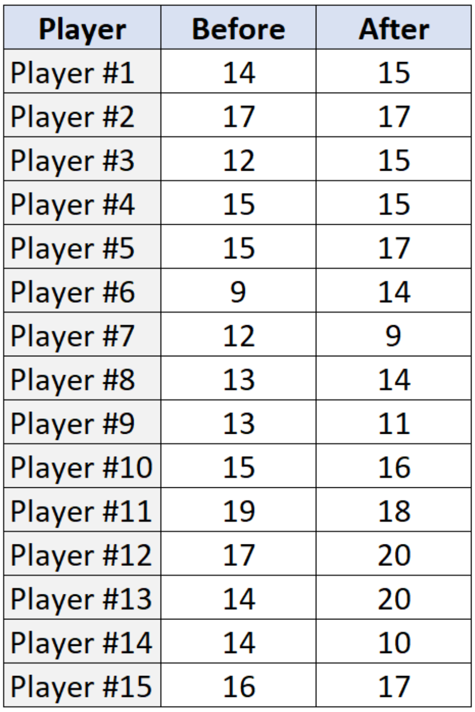
How to Perform the Wilcoxon Signed Rank Test

The following example illustrates how to perform the Wilcoxon Signed Rank test.

> A basketball coach want to know if a certain training program increases the number of free throws made by his players. To

> test this, he has 15 players shoot 20 free throws each before and after the training program.
>
> Since each player can be "paired" with themselves, the coach had planned on using a paired t-test to determine if there was a significant difference between the mean number of free throws made before and after the training program.
>
> However, the distribution of the differences turns out to be non-normal, so the coach instead uses a Wilcoxon Signed Rank Test.

The following table shows the number of free throws made (out of 20 attempts) by each of the 15 players, both before and after the training program:



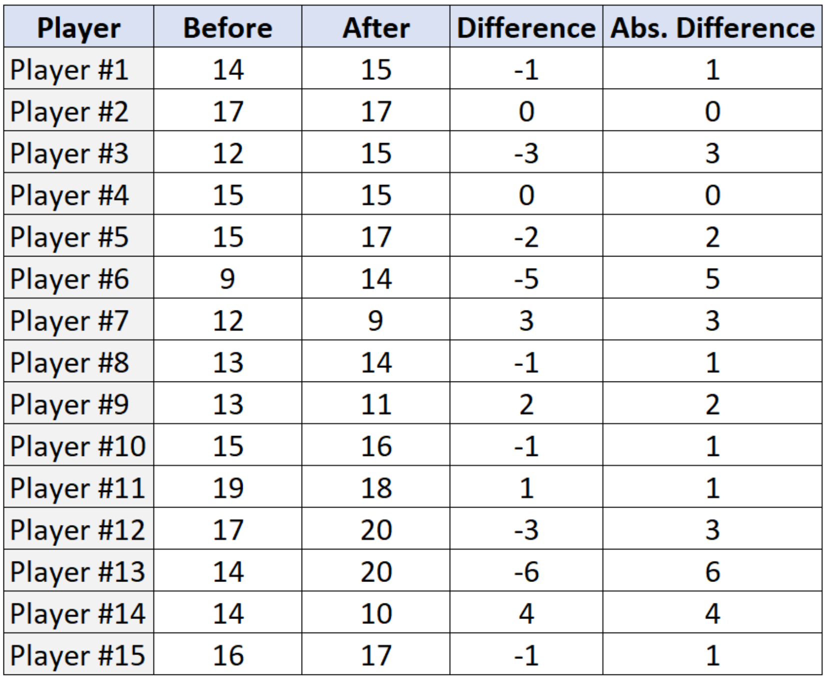
Player	Before	After
Player #1	14	15
Player #2	17	17
Player #3	12	15
Player #4	15	15
Player #5	15	17
Player #6	9	14
Player #7	12	9
Player #8	13	14
Player #9	13	11
Player #10	15	16
Player #11	19	18
Player #12	17	20
Player #13	14	20
Player #14	14	10
Player #15	16	17

Step 1: State the null and alternative hypotheses.

H_0: The median difference between the two groups is zero.

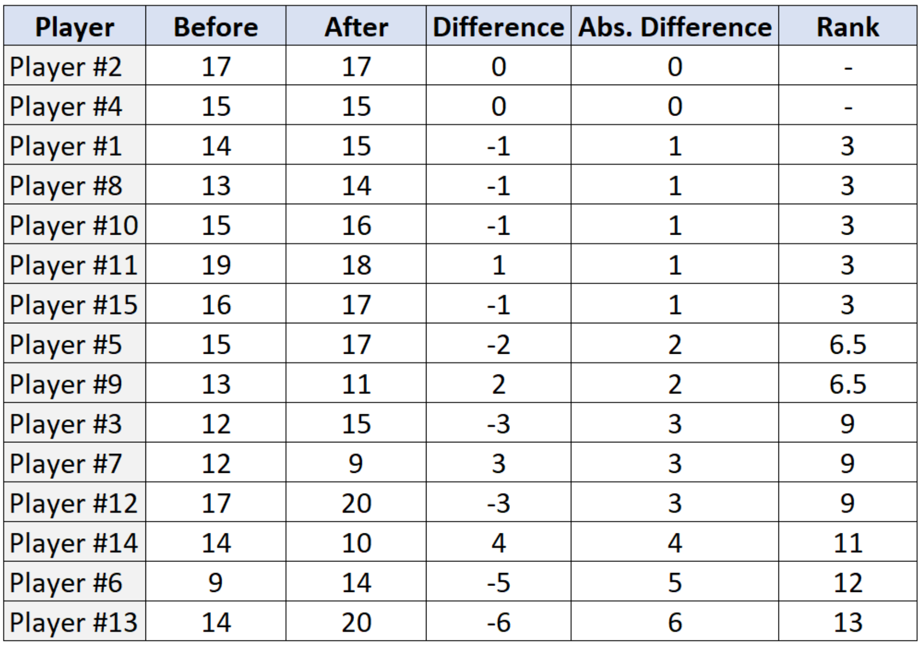
H_A: The median difference is negative. (e.g. the players make less free throws before participating in the training program)

Step 2: Find the difference and absolute difference for each pair.



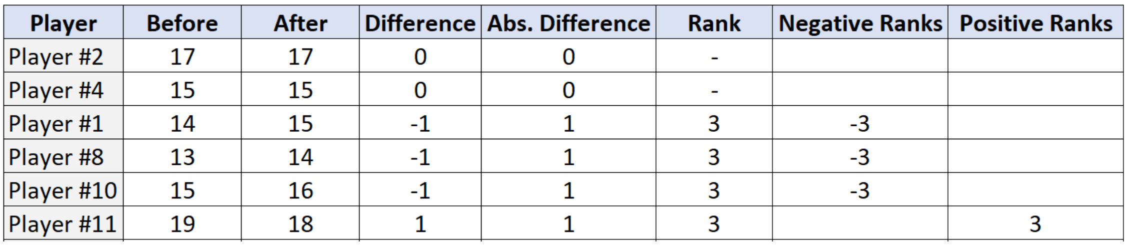
Player	Before	After	Difference	Abs. Difference
Player #1	14	15	-1	1
Player #2	17	17	0	0
Player #3	12	15	-3	3
Player #4	15	15	0	0
Player #5	15	17	-2	2
Player #6	9	14	-5	5
Player #7	12	9	3	3
Player #8	13	14	-1	1
Player #9	13	11	2	2
Player #10	15	16	-1	1
Player #11	19	18	1	1
Player #12	17	20	-3	3
Player #13	14	20	-6	6
Player #14	14	10	4	4
Player #15	16	17	-1	1

Step 3: Order the pairs by the absolute differences and assign a rank from the smallest to largest absolute differences. *Ignore pairs that have an absolute difference of "0" and assign mean ranks when there are ties.*

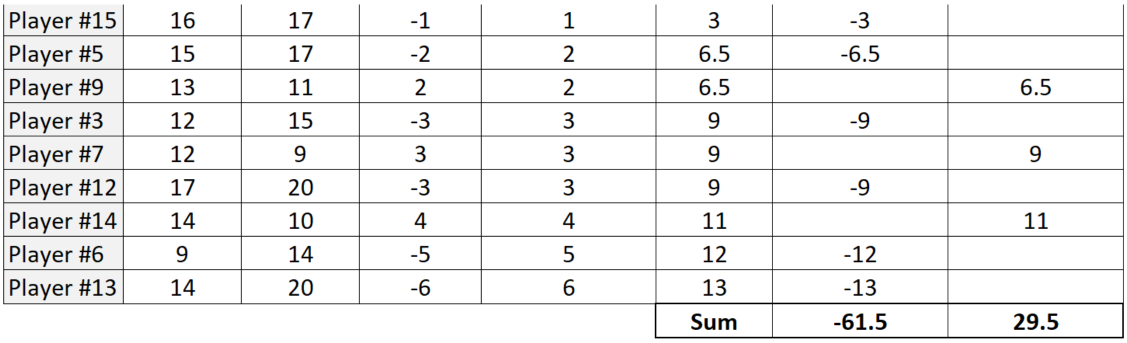


Player	Before	After	Difference	Abs. Difference	Rank
Player #2	17	17	0	0	-
Player #4	15	15	0	0	-
Player #1	14	15	-1	1	3
Player #8	13	14	-1	1	3
Player #10	15	16	-1	1	3
Player #11	19	18	1	1	3
Player #15	16	17	-1	1	3
Player #5	15	17	-2	2	6.5
Player #9	13	11	2	2	6.5
Player #3	12	15	-3	3	9
Player #7	12	9	3	3	9
Player #12	17	20	-3	3	9
Player #14	14	10	4	4	11
Player #6	9	14	-5	5	12
Player #13	14	20	-6	6	13

Step 4: Find the sum of the positive ranks and the negative ranks.



Player	Before	After	Difference	Abs. Difference	Rank	Negative Ranks	Positive Ranks
Player #2	17	17	0	0	-		
Player #4	15	15	0	0	-		
Player #1	14	15	-1	1	3	-3	
Player #8	13	14	-1	1	3	-3	
Player #10	15	16	-1	1	3	-3	
Player #11	19	18	1	1	3		3



Player #15	16	17	-1	1	3	-3	
Player #5	15	17	-2	2	6.5	-6.5	
Player #9	13	11	2	2	6.5		6.5
Player #3	12	15	-3	3	9	-9	
Player #7	12	9	3	3	9		9
Player #12	17	20	-3	3	9	-9	
Player #14	14	10	4	4	11		11
Player #6	9	14	-5	5	12	-12	
Player #13	14	20	-6	6	13	-13	
					Sum	**-61.5**	**29.5**

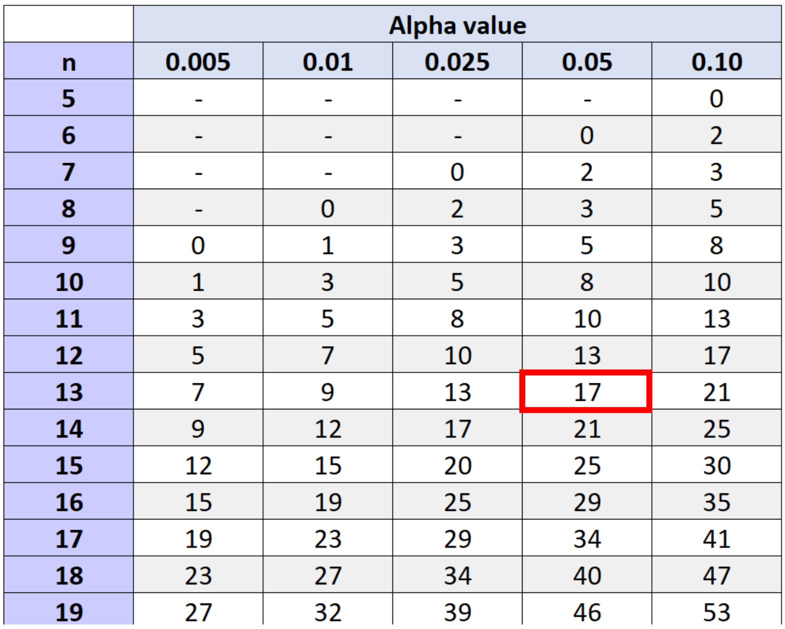
Step 5: Reject or fail to reject the null hypothesis.

The test statistic, W, is the smaller of the absolute values of the positive ranks and negative ranks. In this case, the smaller value is 29.5. Thus, our test statistic is W = **29.5**.

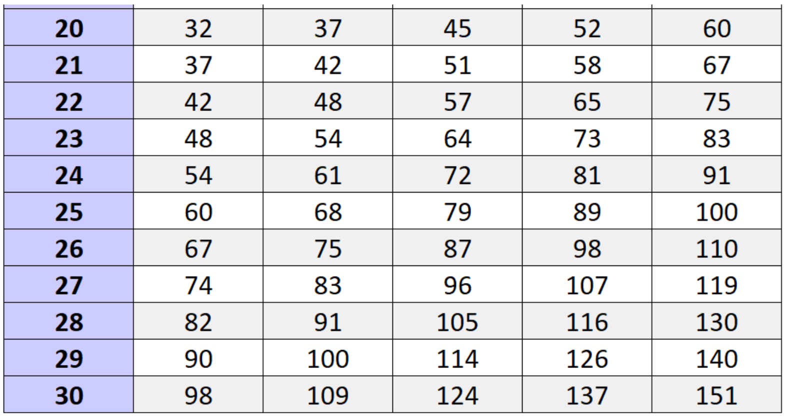
To determine if we should reject or fail to reject the null hypothesis, we can reference the critical value found in the Wilcoxon Signed Rank Test Critical Values Table that corresponds with *n* and our chosen alpha level.

If our test statistic, W, is *less than or equal* to the critical value in the table, we can reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

The critical value that corresponds to an alpha level of 0.05 and n = 13 (the total number of pairs minus the two we didn't calculate ranks for since they had an observed difference of 0) is **17**.



	Alpha value				
n	**0.005**	**0.01**	**0.025**	**0.05**	**0.10**
5	-	-	-	-	0
6	-	-	-	0	2
7	-	-	0	2	3
8	-	0	2	3	5
9	0	1	3	5	8
10	1	3	5	8	10
11	3	5	8	10	13
12	5	7	10	13	17
13	7	9	13	17	21
14	9	12	17	21	25
15	12	15	20	25	30
16	15	19	25	29	35
17	19	23	29	34	41
18	23	27	34	40	47
19	27	32	39	46	53



20	32	37	45	52	60
21	37	42	51	58	67
22	42	48	57	65	75
23	48	54	64	73	83
24	54	61	72	81	91
25	60	68	79	89	100
26	67	75	87	98	110
27	74	83	96	107	119
28	82	91	105	116	130
29	90	100	114	126	140
30	98	109	124	137	151

Since our test statistic (W = 29.5) is not less than or equal to 17, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the training program leads to a significant increase in the number of free throws made by the players.

Note: *Use the Wilcoxon Signed-Rank Test Calculator if you wish to perform the test using a calculator instead of by hand.*

SUS: A quick and dirty usability scale

John Brooke
Contingent Solutions ltd

SUS - A quick and dirty usability scale

John Brooke

Redhatch Consulting Ltd.,
12 Beaconsfield Way,
Earley, READING RG6 2UX
United Kingdom

email: *john.brooke@redhatch.co.uk*

Abstract

Usability does not exist in any absolute sense; it can only be defined with reference to particular contexts. This, in turn, means that there are no absolute measures of usability, since, if the usability of an artefact is defined by the context in which that artefact is used, measures of usability must of necessity be defined by that context too. Despite this, there is a need for broad general measures which can be used to compare usability across a range of contexts. In addition, there is a need for "quick and dirty" methods to allow low cost assessments of usability in industrial systems evaluation. This chapter describes the System Usability Scale (SUS) a reliable, low-cost usability scale that can be used for global assessments of systems usability.

Usability and context

Usability is not a quality that exists in any real or absolute sense. Perhaps it can be best summed up as being a general quality of the ***appropriateness to a purpose*** of any particular artefact. This notion is neatly summed up by Terry Pratchett in his novel "Moving Pictures":

> " 'Well, at least he keeps himself fit,' said the Archchancellor nastily. 'Not like the rest of you fellows. I went into the Uncommon Room this morning and it was full of chaps snoring!'
>
> 'That would be the senior masters, Master,' said the Bursar. 'I would say they are supremely fit, myself.'
>
> '*Fit?* The Dean looks like a man who's swallered a bed!'
>
> 'Ah, but Master,' said the Bursar, smiling indulgently, 'the word "fit",as I understand it, means "appropriate to a purpose", and I would say that the body of the Dean is supremely appropriate to the purpose of sitting around all day and eating big heavy meals.' The Dean permitted himself a little smile. " (Pratchett, 1990)

In just the same way, the usability of any tool or system has to be viewed in terms of the context in which it is used, and its appropriateness to that context. With particular reference to information systems, this view of usability is reflected in the current draft international standard ISO 9241-11 and in the European Community ESPRIT project MUSiC (Measuring Usability of Systems in Context) (e.g., Bevan, Kirakowski and Maissel, 1991). In general, it is impossible to specify the usability of a system (i.e., its fitness for purpose) without first defining who are the intended users of the system, the tasks those users will perform with it, and the characteristics of the physical, organisational and social environment in which it will be used.

Since usability is itself a moveable feast, it follows that measures of usability must themselves be dependent on the way in which usability is defined. It is possible to talk of some general classes of usability measure; ISO 9241-11 suggests that measures of usability should cover

- effectiveness (the ability of users to complete tasks using the system, and the quality of the output of those tasks),
- efficiency (the level of resource consumed in performing tasks)
- satisfaction (users' subjective reactions to using the system).

However, the precise measures to be used within each of these classes of metric can vary widely. For example, measures of effectiveness are very obviously determined by the types of task that are carried out with the system; a measure of effectiveness of a word processing system might be the number of letters written, and whether the letters produced are free of spelling mistakes. If the system supports the task of controlling an industrial process producing chemicals, on the other hand, the measures of task completion and quality are obviously going to reflect that process.

A consequence of the context-specificity of usability and measures of usability is that it is very difficult to make comparisons of usability across different systems. Comparing usability of different systems intended for different purposes is a clear case of "comparing apples and oranges" and should be avoided wherever possible. It is also difficult and potentially misleading to generalise design features and experience across systems; for example, just because a particular design feature has proved to be very useful in making one system usable does not necessarily mean that it will do so for another system with a different group of users doing different tasks in other environments.

If there is an area in which it is possible to make more generalised assessments of usability, which could bear cross-system comparison, it is the area of subjective assessments of usability. Subjective measures of usability are usually obtained through the use of questionnaires and attitude scales, and examples exist of general attitude scales which are not specific to any particular system (for example, CUSI (Kirakowski and Corbett, 1988)).

Industrial usability evaluation

The demands of evaluating usability of systems within an industrial context mean that often it is neither cost-effective nor practical to perform a full-blown context analysis and selection of suitable metrics. Often, all that is needed is a general indication of the overall level of usability of a system compared to its competitors or its predecessors. Equally, when selecting metrics, it is often desirable to have measures which do not require vast effort and expense to collect and analyse data.

These sorts of considerations were very important when, while setting up a usability engineering programme for integrated office systems engineering with Digital Equipment Co. Ltd, a need was identified for a subjective usability measure. The measure had to be capable of being administered quickly and simply, but also had to be reliable enough to be used to make comparisons of user performance changes from version to version of a software product.

The need for simplicity and speed came from the evaluation methods being used; users from customer sites would either visit a human factors laboratory, or a travelling laboratory would be set up at the customer site. The users would then work through evaluation exercises lasting between 20 minutes and an hour, at the end of which a subjective measure of system usability would be collected. As can be imagined, after this period of time, users could be very frustrated, especially if they had encountered problems, since no assistance was given. If they were then presented with a long questionnaire, containing in excess of 25 questions it was very likely that they would not complete it and there would be insufficient data to assess subjective reactions to system usability.

SUS - the System Usability Scale

In response to these requirements, a simple usability scale was developed. The System Usability Scale (SUS) is a simple, ten-item scale giving a global view of subjective assessments of usability.

SUS is a *Likert scale*. It is often assumed that a Likert scale is simply one based on forced-choice questions, where a statement is made and the respondent then indicates the degree of agreement or disagreement with the statement on a 5 (or 7) point scale. However, the construction of a Likert scale is somewhat more subtle than this. Whilst Likert scales are presented in this form, the statements with which the respondent indicates agreement and disagreement have to be selected carefully.

The technique used for selecting items for a Likert scale is to identify examples of things which lead to extreme expressions of the attitude being captured. For instance, if one was interested in attitudes to crimes and misdemeanours, one might use serial murder and parking offences as examples of the extreme ends of the spectrum. When these examples have been selected, then a sample of respondents is asked to give ratings to these examples across a wide pool of potential questionnaire items. For instance, respondents might be asked to respond to statements such as “hanging’s too good for them”, or “I can imagine myself doing something like this”.

Given a large pool of such statements, there will generally be some where there is a lot of agreement between respondents. In addition, some of these will be ones where the statements provoke extreme statements of agreement or disagreement among all respondents. It is these latter statements which one tries to identify for inclusion in a Likert scale, since, we would hope that, if we have selected suitable examples, there would be general agreement of extreme attitudes to them. Items where there is ambiguity are not good discriminators of attitudes. For instance, while one hopes that there would be a general, extreme disagreement that “hanging’s too good” for those who perpetrate parking offences, there may well be less agreement about applying this statement to serial killers, since opinions differ widely about the ethics and efficacy of capital punishment.

SUS was constructed using this technique. A pool of 50 potential questionnaire items was assembled. Two examples of software systems were then selected (one a linguistic tool aimed at end users, the other a tool for systems programmers) on the basis of general agreement that one was “really easy to use” and one was almost impossible to use, even for highly technically skilled users. 20 people from the office systems engineering group, with occupations ranging from secretary through to systems programmer then rated both systems against all 50 potential questionnaire items on a 5 point scale ranging from “strongly agree” to “strongly disagree”.

The items leading to the most extreme responses from the original pool were then selected. There were very close intercorrelations between all of the selected items (± 0.7 to ± 0.9). In addition, items were selected so that the common response to half of them was strong agreement, and to the other half, strong disagreement. This was done in order to prevent response biases caused by respondents not having to think about each statement; by alternating positive and negative items, the respondent has to read each statement and make an effort to think whether they agree or disagree with it.

The System Usability Scale is shown in the next section of this chapter. It can be seen that the selected statements actually cover a variety of aspects of system usability, such as the need for support, training, and complexity, and thus have a high level of face validity for measuring usability of a system.

System Usability Scale

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	□ 1	□ 2	□ 3	□ 4	□ 5
2. I found the system unnecessarily complex	□ 1	□ 2	□ 3	□ 4	□ 5
3. I thought the system was easy to use	□ 1	□ 2	□ 3	□ 4	□ 5
4. I think that I would need the support of a technical person to be able to use this system	□ 1	□ 2	□ 3	□ 4	□ 5
5. I found the various functions in this system were well integrated	□ 1	□ 2	□ 3	□ 4	□ 5
6. I thought there was too much inconsistency in this system	□ 1	□ 2	□ 3	□ 4	□ 5
7. I would imagine that most people would learn to use this system very quickly	□ 1	□ 2	□ 3	□ 4	□ 5
8. I found the system very cumbersome to use	□ 1	□ 2	□ 3	□ 4	□ 5
9. I felt very confident using the system	□ 1	□ 2	□ 3	□ 4	□ 5
10. I needed to learn a lot of things before I could get going with this system	□ 1	□ 2	□ 3	□ 4	□ 5

Using SUS

The SU scale is generally used after the respondent has had an opportunity to use the system being evaluated, but before any debriefing or discussion takes place. Respondents should be asked to record their immediate response to each item, rather than thinking about items for a long time.

All items should be checked. If a respondent feels that they cannot respond to a particular item, they should mark the centre point of the scale.

Scoring SUS

SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own.

To calculate the SUS score, first sum the score contributions from each item. Each item's score contribution will range from 0 to 4. For items 1,3,5,7,and 9 the score contribution is the scale position minus 1. For items 2,4,6,8 and 10, the contribution is 5 minus the scale position. Multiply the sum of the scores by 2.5 to obtain the overall value of SU.

SUS scores have a range of 0 to 100.

The following section gives an example of a scored SU scale.

System Usability Scale

	Strongly disagree 1	2	3	4	Strongly agree 5	
1. I think that I would like to use this system frequently					√	4
2. I found the system unnecessarily complex				√		1
3. I thought the system was easy to use		√				1
4. I think that I would need the support of a technical person to be able to use this system	√					4
5. I found the various functions in this system were well integrated		√				1
6. I thought there was too much inconsistency in this system			√			2
7. I would imagine that most people would learn to use this system very quickly		√				1
8. I found the system very cumbersome to use				√		1
9. I felt very confident using the system					√	4
10. I needed to learn a lot of things before I could get going with this system		√				3

Total score = 22

SUS Score = 22 *2.5 = 55

Conclusion

SUS has proved to be a valuable evaluation tool, being robust and reliable. It correlates well with other subjectives measures of usability (eg., the general usability subscale of the SUMI inventory developed in the MUSiC project (Kirakowski, personal communication)). SUS has been made freely available for use in usability assessment, and has been used for a variety of research projects and industrial evaluations; the only prerequisite for its use is that any published report should acknowledge the source of the measure.

Acknowledgements

SUS was developed as part of the usability engineering programme in integrated office systems development at Digital Equipment Co Ltd., Reading, United Kingdom.

References

Bevan, N, Kirakowski, J and Maissel, J, 1991, What is Usability?, in H.-J. Bullinger, (Ed.). *Human Aspects in Computing: Design and use of interactive systems and work with terminals*, Amsterdam: Elsevier.

Kirakowski, J and Corbett, M, 1988, Measuring User Satisfaction, in D M Jones and R Winder (Eds.) *People and Computers IV*. Cambridge: Cambridge University Press.

Pratchett, T., 1990 *Moving Pictures*. London: Gollancz

Code Inspection Checklist (Java)

Variable and constant declarations

1. Are descriptive variable and constant names used in accordance with naming conventions?

   Descriptive names make code easier to read, understand, and maintain.

2. Are there variables with confusingly similar names?

   If this is the case, it would be easy to type one name when you meant another. If the variables are of the same type, the compiler won't detect your mistake.

3. Is every variable *correctly* initialised?

   The compiler requires that all variables are initialised, but are they initialised to the right value?

4. Could any non-local variables be made local?

   If variables with class scope (i.e. instance variables) are only used by a single method and this method doesn't have to remember its value between calls, the variable can be made local to the method. Unnecessary non-local variables make classes overly complex and thus more difficult to read, understand, and maintain.

5. Are there literal constants that should be named constants?

   Replace occurrences of literal numbers, like 0.175 for the VAT rate, with named constants (e.g. `VAT_RATE` *where this is declared as* `final int VAT_RATE = 0.175`*). Should the VAT rate change, you only need to change the value of the constant in your program, and not every occurrence of 0.175. Using a named constant, you can be sure that the right value is being used in all cases.*

6. Are there variables that should be constants?

   If a class has a variable whose value isn't intended to change, you should make it a constant so that it can never be changed accidentally. You can then rely on the compiler to detect an attempted modification.

Methods

7. Are descriptive method names used in accordance with naming conventions?

Descriptive names make code easier to read, understand, and maintain.

8. Is every method parameter (argument) checked before being used?

   It is good practice for methods to check their argument's values before using them. E.g. when setting a name, check for a null String and don't change the object's state if the argument is null.

9. For every method, does it return the correct value at every method return point?

Operators

10. For each expression with more than one operator, are the assumptions about order of evaluation and precedence correct?

    Use brackets to clear up ambiguities.

11. Are comparison operators (i.e. < <= > >= ==) correct.

12. Is each boolean expression correct?

In general, use the equality operator (and not the assignment operator) when writing boolean expressions.

Control flow (if, switch, for while)

13. For each loop, is the most appropriate construct used?

    Use while for loops where you don't know in advance how many iterations are to be made; use for loops in other cases.

14. Will all loops terminate?

15. Where there are multiple exists from a loop, is each exit point necessary and handled correctly?

16. Does each switch statement have a default case?

    In general, provide a default label for every switch statement.

17. Are missing break statements (in a switch construct) correct and marked with a comment?

    In general follow cases statements with a break.

18. Do any if structures have dangling else clauses?

    Avoid dangling else statements.

Interfaces

19. Is the order of arguments in every method call in agreement with the method's declaration?

    Where a method has two arguments of the same type, for example, the compiler can't detect when you call the method that the actual arguments are given in the correct order.

20. Do the values in units agree?

    Recall the Mars Climate Orbiter system mentioned in lecture 1. One component supplied Metric values to another that interpreated them (without any conversion) as Imperial units.

Comments

21. Does every class and method have an appropriate header comment?

22. Does every variable or constant have a comment?

23. Is the underlying behaviour of each class and method expressed in plain English?

24. Do the comments and code agree?

    Comments that describe how the code used to work are of little value once the code has been changed.

25. Do the comments help you to understand the code?

    Comments should add value to the source code.

Input/output

26. Are the spelling or grammatical errors in any text printed or displayed?

27. Are erroneous input data checked?

Acknowledgement

This checklist is based on the Java Code Inspection Checklist from Ian Warren.

Black Box Testing: Equivalence Partitioning

Introduction

Equivalence partitioning is based on the premise that the inputs and outputs of a component can be partitioned into classes that, according to the component's specification, will be treated similarly by the component. Thus the result of testing a single value from an equivalence partition is considered representative of the complete partition. **Example**

Consider a component, *generate_grading,* with the following specification:

> *The component is passed an exam mark (out of 75) and a coursework (c/w) mark (out of 25), from which it generates a grade for the course in the range 'A' to 'D'. The grade is calculated from the overall mark which is calculated as the sum of the exam and c/w marks, as follows:*
>
> | | | |
> |---|---|---|
> | *greater than or equal to 70* | - | *'A'* |
> | *greater than or equal to 50, but less than 70* | - | *'B'* |
> | *greater than or equal to 30, but less than 50* | - | *'C'* |
> | *less than 30* | - | *'D'* |
>
> *Where a mark is outside its expected range then a fault message ('FM') is generated. All inputs are passed as integers.*

Initially the equivalence partitions are identified and then test cases derived to exercise the partitions. Equivalence partitions are identified from both the inputs and outputs of the component and both valid and invalid inputs and outputs are considered.

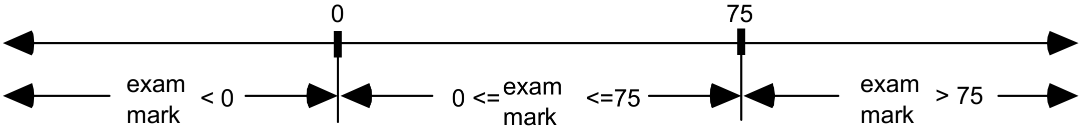
The partitions for the two inputs are initially identified. The *valid* partitions can be described by:

$0 \leq$ exam mark ≤ 75
$0 \leq$ coursework mark ≤ 25

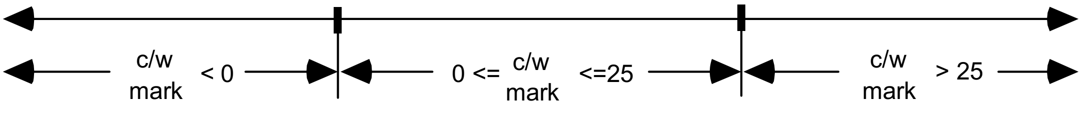
The most obvious *invalid* partitions based on the inputs can be described by:

exam mark > 75 exam
mark < 0 coursework
mark > 25
coursework mark < 0

Partitioned ranges of values can be represented pictorially, therefore, for the input, exam mark, we get:



And for the input, coursework mark, we get:



Less obvious invalid input equivalence partitions would include any other inputs that can occur not so far included in a partition, for instance, non-integer inputs or perhaps non-numeric inputs. So, we could generate the following invalid input equivalence partitions:

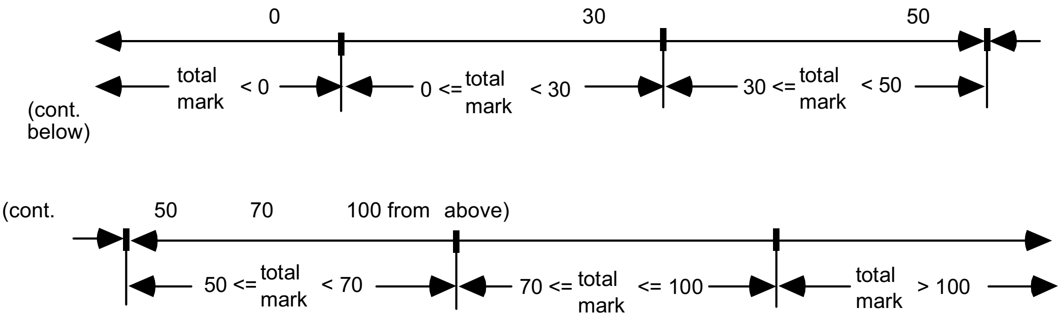
exam mark = real number (a number with a fractional part)
exam mark = alphabetic
coursework mark = real number
coursework mark = alphabetic

Next, the partitions for the outputs are identified. The *valid* partitions are produced by considering each of the valid outputs for the component:

'A'	is induced by	$70 \le$ total mark ≤ 100
'B'	is induced by	$50 \le$ total mark < 70
'C'	is induced by	$30 \le$ total mark < 50
'D'	is induced by	$0 \le$ total mark < 30
'Fault Message'	is induced by	total mark > 100
'Fault Message'	is induced by	total mark < 0

where total mark = exam mark + coursework mark. Note that 'Fault Message' is considered as a valid output as it is a *specified* output.

The equivalence partitions and boundaries for total mark are shown pictorially below:



An invalid output would be any output from the component other than one of the five specified. It is difficult to identify unspecified outputs, but obviously they must be considered as if we can cause one then we have identified a flaw with either the component, its specification, or both. For this example three unspecified outputs were identified and are shown below. This aspect of equivalence partitioning is very subjective and different testers will inevitably identify different partitions which *they* feel could possibly occur.

output	= 'E'
output	= 'A+'
output	= 'null'

Thus the following nineteen equivalence partitions have been identified for the component (remembering that for some of these partitions a certain degree of subjective choice was required, and so a different tester would not necessarily duplicate this list exactly):

$0 \le$ exam mark ≤ 75
exam mark > 75
exam mark < 0
$0 \le$ coursework mark ≤ 25
coursework mark > 25
coursework mark < 0
exam mark = real number

exam mark = alphabetic
coursework mark = real number
coursework mark = alphabetic
70 ≤ total mark ≤ 100
50 ≤ total mark < 70
30 ≤ total mark < 50
0 ≤ total mark < 30
total mark > 100
total mark < 0
output = 'E'
output = 'A+'
output = 'null'

Having identified all the partitions then test cases are derived that 'hit' each of them. Two distinct approaches can be taken when generating the test cases. In the first a test case is generated for each identified partition on a one-to-one basis, while in the second a minimal set of test cases is generated that cover all the identified partitions.

The one-to-one approach will be demonstrated first as it can make it easier to see the link between partitions and test cases. For each of these test cases only the single partition being targetted is stated explicitly. Nineteen partitions were identified leading to nineteen test cases.

The test cases corresponding to partitions derived from the input exam mark are:

Test Case	1	2	3
Input (exam mark)	44	-10	93
Input (c/w mark)	15	15	15
total mark (as calculated)	59	5	108
Partition tested (of exam mark)	0 ≤ e ≤ 75	e < 0	e > 75
Exp. Output	'B'	'FM'	'FM'

Note that the input coursework (c/w) mark has been set to an arbitrary valid value of 15.

The test cases corresponding to partitions derived from the input coursework mark are:

Test Case	4	5	6
Input (exam mark)	40	40	40
Input (c/w mark)	8	-15	47
total mark (as calculated)	48	25	87
Partition tested (of c/w mark)	0 ≤ c ≤ 25	c < 0	c > 25
Exp. Output	'C'	'FM'	'FM'

Note that the input exam mark has been set to an arbitrary valid value of 40.

The test cases corresponding to partitions derived from possible invalid inputs are:

Test Case	7	8	9	10
Input (exam mark)	48.7	q	40	40
Input (c/w mark)	15	15	12.76	g
total mark (as calculated)	63.7	not applicable	52.76	not applicable
Partition tested	exam mark = real number	exam mark = alphabetic	c/w mark = real number	c/w mark = alphabetic
Exp. Output	'FM'	'FM'	'FM'	'FM'

The test cases corresponding to partitions derived from the valid outputs are:

Test Case	11	12	13
Input (exam mark)	-10	12	32
Input (c/w mark)	-10	5	13
total mark (as calculated)	-20	17	45
Partition tested (of total mark)	t < 0	0 ≤ t < 30	30 ≤ t < 50
Exp. Output	'FM'	'D'	'C'

Test Case	14	15	16
Input (exam mark)	44	60	80
Input (c/w mark)	22	20	30
total mark (as calculated)	66	80	110
Partition tested (of total mark)	50 ≤ t < 70	70 ≤ t ≤ 100	t > 100
Exp. Output	'B'	'A'	'FM'

The input values of exam & coursework marks have been derived from the total mark, which is their sum.

The test cases corresponding to partitions derived from the invalid outputs are:

Test Case	17	18	19
Input (exam mark)	-10	100	null
Input (c/w mark)	0	10	null
total mark (as calculated)	-10	110	null+null
Partition tested (output)	'E'	'A+'	'null'
Exp. Output	'FM'	'FM'	'FM'

It should be noted that where invalid input values are used (as above, in test cases 2, 3, 5-11, and 1619) it may, depending on the implementation, be impossible to actually execute the test case. For instance, in Ada, if the input variable is declared as a positive integer then it will not be possible to assign a negative value to it. Despite this, it is still worthwhile *considering* all the test cases for completeness.

It can be seen above that several of the test cases are similar, such as test cases 1 and 14, where the main difference between them is the partition targetted. As the component has two inputs and one output, each test case actually 'hits' three partitions; two input partitions and one output partition.. Thus it is possible to generate a smaller 'minimal' test set that still 'hits' all the identified partitions by deriving test cases that are designed to exercise more than one partition. The following test case suite of eleven test cases corresponds to the minimised test case suite approach where each test case is designed to hit as many new partitions as possible rather than just one. Note that here all three partitions are explicitly identified for each test case.

Test Case	1	2	3	4
Input (exam mark)	60	40	25	15
Input (c/w mark)	20	15	10	8
total mark (as calculated)	80	55	35	23
Partition (of exam mark)	$0 \leq e \leq 75$	$0 \leq e \leq 75$	$0 \leq e \leq 75$	$0 \leq e \leq 75$
Partition (of c/w mark)	$0 \leq c \leq 25$	$0 \leq c \leq 25$	$0 \leq c \leq 25$	$0 \leq c \leq 25$
Partition (of total mark)	$70 \leq t \leq 100$	$50 \leq t < 70$	$30 \leq t < 50$	$0 \leq t < 30$
Exp. Output	'A'	'B'	'C'	'D'

Test Case	5	6	7	8
Input (exam mark)	-10	93	60.5	q
Input (c/w mark)	-15	35	20.23	g
total mark (as calculated)	-25	128	80.73	-
Partition (of exam mark)	$e < 0$	$e > 75$	e = real number	e = alphabetic
Partition (of c/w mark)	$c < 0$	$c > 25$	c = real number	c = alphabetic
Partition (of total mark)	$t < 0$	$t > 100$	$70 \leq t \leq 100$	-
Exp. Output	'FM'	'FM'	'FM'	'FM'

Test Case	9	10	11
Input (exam mark)	-10	100	'null'
Input (c/w mark)	0	10	'null'
total mark (as calculated)	-10	110	null+null
Partition (of exam mark)	$e < 0$	$e > 75$	-
Partition (of c/w mark)	$0 \leq c \leq 25$	$0 \leq c \leq 25$	-
Partition (of total mark)	$t < 0$	$t > 100$	-
Partition (of output)	'E'	'A+'	'null'
Exp. Output	'FM'	'FM'	'FM'

The one-to-one and minimised approaches represent the two approaches to equivalence partitioning. The disadvantage of the one-to-one approach is that it requires more test cases and if this causes problems a more minimalist approach can be used. Normally, however, the identification of partitions is far more time consuming than the generation and execution of test cases themselves and so any savings made by reducing the size of the test case suite are relatively small compared with the overall cost of applying the technique. The disadvantage of the minimalist approach is that in the event of a test failure it can be difficult to identify the cause due to several new partitions being exercised at once. This is a debugging problem rather than a testing problem, but there is no reason to make debugging more difficult than it is already.

Some testers would say that a variable's input domain equivalence partitions must be combined with every other variables' input domains equivalence partitions. This is an extreme view as it leads to an explosion of number of tests. It hard however to argue against it on the ground of completeness.