NASA TASK LOAD INDEX (TLX)

v. 1.0

Paper and Pencil Package

Human Performance Research Group
NASA Ames Research Center
Moffett Field, California
(415)694-6072

Table of Contents

1. Background 1
2. Description 2
   2.1 General Information
   2.2 Sources of Load (WEIGHTS)
   2.3 Magnitude of Load (RATINGS)
   2.4 Weighting and Averaging Procedure (COMBINE)
3. Experimental Procedure 4
   3.1 Overview of Programs Provided
   3.2 Computer Setup
   3.3 Data Input for RATINGS and WEIGHTS Programs
      3.3.1 General Description
      3.3.2 Selecting Condition Labels
      3.3.3 How Weights are Applied to Ratings
      3.3.4 How to Input Data
   3.4 Collecting Subject Ratings
      3.4.1 Instructions
      3.4.2 Familiarization
      3.4.3 Using the Program
   3.5 Collecting Subject Weights
   3.6 Summary
4. Data Analysis Procedure 10
   4.1 Using the Program
5. Bibliography 12
6. Subject Instructions: Ratings (Keyboard Version) 15
7. Subject Instructions: Sources-of-Workload Evaluation (Keyboard Version) 16
8. Subject Instructions: Ratings (Mouse Version) 17
9. Subject Instructions: Sources-of-Workload Evaluation (Mouse Version) 18

Appendix A. Rating Scale Definitions 19
Appendix B. Rating Scales 20

NASA Task Load Index (NASA-TLX)
Version 1.0

Computerized Version

This booklet and the accompanying diskette contain the materials necessary to collect subjective workload assessments with the NASA Task Load Index on IBM PC compatible microcomputers. This procedure for collecting workload ratings was developed by the Human Performance Group at NASA Ames Research Center during a three year research effort that involved more than 40 laboratory, simulation, and inflight experiments. Although the technique is still undergoing evaluation, this package is being distributed to allow other researchers to use it in their own experiments. Comments or suggestions about the procedure would be greatly appreciated. This package is intended to fill a "nuts and bolts" function of describing the procedure. A bibliography provides background information about previous empirical findings and the logic that supports the procedure.

1. BACKGROUND

The NASA Task Load Index is a multi-dimensional rating procedure that provides an overall workload score based on a weighted average of ratings on six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort, and Frustration. A definition of each subscale is provided in Appendix A.

An earlier version of the scale had nine subscales. It was designed to reduce between-rater variability by using the *a priori* workload definitions of subjects to weight and average subscale ratings. This technique (referred to as the "NASA Bipolar Rating Scale") was quite successful in reducing between-rater variability, and it provided diagnostic information about the magnitudes of different sources of load from subscale ratings (Hart, Battiste, & Lester, 1984; Vidulich & Tsang, 1985a & b). However, its sensitivity to experimental manipulations, while better than found for other popular techniques and for a global unidimensional workload rating, was still not considered sufficient. In addition, it was felt that nine subscales are too many, making the scale impractical to use in a simulation or operational environment. Finally, several of the subscales were found to be irrelevant to workload (e.g., Fatigue) or redundant (e.g., Stress and Frustration). For these reasons, the NASA Task Load Index was developed. Some of the

subscales from the original scale were revised or combined, others deleted, and two added. Three dimensions relate to the demands imposed on the subject (Mental, Physical, and Temporal Demands) and three to the interaction of a subject with the task (Effort, Frustration, and Performance).

Although it is clear that definitions of workload do indeed vary among experimenters and among subjects (contributing to confusion in the workload literature and between-rater variability), it was found that the specific sources of loading imposed by different tasks are an even more important determinant of workload experiences. Thus, the current version of the scale (the Task Load Index) combines subscale ratings that are weighted according to their subjective importance to raters in a specific task, rather than their *a priori* relevance to raters' definitions of workload in general.

2. DESCRIPTION

2.1. General Information

The degree to which each of the six factors contribute to the workload of the specific task to be evaluated, from the raters' perspectives, is determined by their responses to pair-wise comparisons among the six factors. Magnitude ratings on each subscale are obtained after each performance of a task or task segment. Ratings of factors deemed most important in creating the workload of a task are given more weight in computing the overall workload score, thereby enhancing the sensitivity of the scale.

The weights and ratings may or may not covary. For example, it is possible for mental demands to be the primary source of loading for a task, even though the magnitude of the mental demands might be low. Conversely, the time pressure under which a task is performed might be the primary source of its workload, and the time demands might be rated as being high for some versions of the task and low for others.

Since subjects can give ratings quickly, it may be possible to obtain them in operational settings. However, a videotaped replay or computer regeneration of the operator's activities may be presented as a mnemonic aid that can be stopped after each segment to obtain ratings retospectively. It was shown in a helicopter simulation and in a supervisory control simulation (Hart, Battiste, Chesney, Ward, & McElroy, 1986; Haworth, Bivens, and Shively, 1986) that little information was lost when ratings were given retrospectively; a high correlation was found between ratings that were obtained "online" and those that were obtained retrospectively with a visual

re-creation of the task.

The Task Load Index has been tested in a variety of experimental tasks that range from simulated flight to supervisory control simulations and laboratory tasks (e.g., the Sternberg memory task, choice reaction time, critical instability tracking, compensatory tracking, mental arithmetic, mental rotation, target acquisition, grammatical reasoning, etc.). The results of the first validation study are summarized in Hart & Staveland (in press). The derived workload scores have been found to have substantially less between-rater variability than unidimensional workload ratings, and the subscales provide diagnostic information about the sources of load.

2.2. Sources of Load (WEIGHTS)

The NASA Task Load Index is a two-part evaluation procedure consisting of both weights and ratings. Three separate computer programs are provided: "WEIGHTS" is used to collect weights; "RATINGS" is used to collect ratings; and "COMBINE" is used to combine them into an overall weighted workload score. The first requirement is for each rater to evaluate the contribution of each factor (its weight) to the workload of a specific task. These weights account for two potential sources of between-rater variability: differences in workload definition between raters within a task, and differences in the sources of workload between tasks. In addition, the weights themselves provide diagnostic information about the nature of the workload imposed by the task.

There are 15 possible pair-wise comparisons of the six scales. The WEIGHTS program presents each pair to the subject on the CRT. Subjects select the member of each pair that contributed more to the workload of that task. The computer tallies the number of times that each factor was selected. The tallies can range from 0 (not relevant) to 5 (more important than any other factor).

A different set of weights is obtained for each distinctly different task or task element upon its completion. The same set of weights can be used for many different versions of the same task if the contributions of the six factors to their workload is fairly similar. For example, the same set of weights was used for many different versions of a target acquisition task in which time pressure, target acquisition difficulty, and decision making load were varied. Obtaining separate weights for different experimental manipulations increased the sensitivity of the derived workload score only slightly, and did not warrant the additional time required to gather them. On the

other hand, the weights obtained from the same subjects for a compensatory tracking task or a memory search task would not have been appropriate for the target acquisition task.

2.3. Magnitude of Load (RATINGS)

The second requirement is to obtain numerical ratings for each scale that reflect the magnitude of that factor in a given task. The RATINGS program presents the six scales on the CRT. Subjects respond by marking each scale at the desired location, using either the keyboard or a mouse. Each scale is presented as a line divided into 20 equal intervals anchored by bipolar descriptors (e.g., High/Low). Ratings may be obtained either during a task, after task segments, or following an entire task. In operational situations, rating sheets or verbal responses are more practical, while the computerized version is more efficient for most simulation and laboratory settings. (A paper and pencil package, for use where a computer is impractical, is available from NASA Ames Research Center.)

2.4. Weighting and Averaging Procedure (COMBINE)

The COMBINE program computes the overall workload score for each subject by multiplying each rating by the weight given to that factor by that subject. The sum of the weighted ratings for each task is divided by 15 (the sum of the weights). The user instructs the program whether separate weights were collected for different subjects, experimental conditions, and replications.

3. EXPERIMENTAL PROCEDURE

3.1. Overview of Programs Provided

The accompanying diskette contains the source code (.PAS) in Turbo Pascal and the executable code (.COM) versions of three programs: RATINGS, WEIGHTS, and COMBINE. Two versions of RATINGS and WEIGHTS are supplied: MRATINGS and MWEIGHTS are for use when subject input is gathered with a mouse, and KRATINGS and KWEIGHTS are for use when subject input is gathered with a keyboard. (Both versions use a keyboard for experimenter input.) Before using the programs, rename the appropriate executable programs "RATINGS" and "WEIGHTS" (e.g., if you intend to use a mouse, rename "MRATINGS.COM" to "RATINGS.COM," etc.). One data file, DEFAULTS, is also supplied, which

contains information used by the program, and must be in the same directory as the program. The source code is provided if you wish to modify the program for your setup. To recompile, you will need Turbo Pascal v. 3.0. Before using the programs, make backup copies.

Three types of data files will be created: RATINGS produces .RAT files, containing the subjects' ratings on the scales; WEIGHTS produces .WGT files, containing the weights obtained during the sources-of-workload evaluation; and COMBINE produces .WWL files, containing the original ratings plus the weighted workload scores. Most researchers will only use the .WWL files. However, interested experimenters may also examine the .WGT files.

3.2. Computer Setup

The program runs on IBM PC compatible microcomputers. The RATINGS program requires an IBM color graphics card (or compatible) to display the scales. The text may be fairly illegible on certain non-RGB monitors. The "mouse" version of the program is written for a Microsoft mouse.

It is desirable to have separate monitors for the experimenter and the subject, with the ability to switch the output to the subject's CRT on and off, so that only the experimenter sees the initial part of the RATINGS and WEIGHTS programs (which require input of setup parameters).

To run the programs, follow the usual procedure for running an executable program. In general, this involves turning the computer on with the operating system installed, then putting the diskette in a floppy disk drive, switching to that drive, and typing th name of the desired program.

Any program can be aborted at any time by typing CTRL-C. However, this may leave an output data file with 0 blocks. This should be deleted. All experimenter input is followed by a carriage return (CR).

3.3. Data Input for the RATINGS and WEIGHTS Programs

The first part of both programs requires the experimenter to enter information to define the experimental conditions for which ratings or weights are being collected. The second part presents the scales (RATINGS) or scale titles (WEIGHTS) to the subject and collects subject responses. As noted above, subjects should see only the second part.

3.3.1. General Description

Both programs need four pieces of information every time ratings or weights are collected from a subject:

(1) The name of the data file in which the ratings or weights will be stored (up to eight or ten characters, depending on your operating system). The extensions (.RAT for ratings, .WGT for weights) will be be supplied by the program. One file can contain ratings or weights from any number of subjects, conditions, and/or replications. Each set of ratings or weights will be stored on a separate line, identified by subject number, condition code, and replication number. The ratings will consist of a number between 0 and 100 for each of the six subscales; the weights will consist of a number between 0 and 1 for each subscale.

(2) A one or two digit subject number (no letters).

(3) A three-character condition code or label (e.g., EX1).

(4) A one or two digit replication number.

The DEFAULTS file contains the filename, subject number, condition code, and replication number most recently used. (The file comes supplied with default values before you use the program, but once you have supplied values, they remain in the file until it is changed. Users should not modify the DEFAULT file themselves; the program takes care of it.) The RATINGS or WEIGHTS programs will ask you if you want to reuse the information you last supplied (e.g., use the same filename, subject number, etc.). If you do, you just type a CR. If not, type in the new information. In most experimental situations, you would enter a filename and subject number at the beginning of a session, and then for multiple runs with the same subject, would only change the condition code and/or replication, keeping the same filename and subject number. The program will not allow you to reuse the same combination of filename, subject number, condition, *and* replication twice; some new identification must be supplied for each trial.

If you realize you have made a mistake in entering data, you can correct errors on the same line, or bail out by typing CTRL-C.

3.3.2. Selecting Condition Labels

Some thought should be given to the names given to experimental conditions, depending on whether you collect separate weights for different

trials. This is because the COMBINE program needs to know whether weights are conditional on condition, and if so, which weights to apply to which sets of ratings. It does this by matching the condition codes supplied for the RATINGS and WEIGHTS programs. Three situations may pertain:

(1) Only one set of weights is collected for all experimental conditions. (For example, you may present a number of tracking tasks which differ so slightly that only one sources-of-workload evaluation is needed.) In this case, you can call the conditions whatever you like, and tell the COMBINE program that weights are *not* conditional on experimental tasks.

(2) A separate set of weights is collected for each experimental condition or task. For example, each subject may do one kind of tracking task and one kind of memory task, and a separate set of weights is collected for each type of task. In this case, supply the same two condition codes (one for tracking, one for memory) when running both RATINGS and WEIGHTS, and tell the COMBINE program that weights *are* conditional on tasks.

(3) There are a number of different experimental conditions, with a smaller set of weights, some applied to certain tasks and some to others. For example, there might be several tracking tasks and several memory tasks, with one set of tracking weights and one set of memory weights. In this case, you have two options:

(a) In the .RAT file, use the same label for all experimental tasks that will use the same weights (i.e., give the same label to all tracking tasks, and a different label to all memory tasks).

(b) If you wish to store distinct labels for different tracking and memory conditions, use different data files: store the tracking ratings in one file and the memory ratings in another file. Then you can use the tracking weights with the tracking ratings file and inform the COMBINE program that weights are *not* conditional on experimental tasks. You can do likewise for memory conditions.

3.3.3. How Weights are Applied to Ratings

If weights *are* conditional on a variable, they are only applied to ratings whose label for that variable exactly matches that for the weight. If weights are *not* conditional on a variable, that variable's label is ignored. For example, if you say weights *are* conditional on subjects, and you supply

a file with *weights* for subjects 1 and 2, and *ratings* for subjects 1, 2, and 3, the appropriate weights will be applied to subjects 1 and 2, and *equal* weights will be applied to subject 3. If you say weights are *not* conditional on subjects, the first set of weights encountered in the .WGT file (presumably, the weights for subject 1) will be applied to all three subjects.

3.3.4. How to Input Data

The first part of both RATINGS and WEIGHTS operates the same way:

(1) Start the program by typing "Ratings" or "Weights," and then respond to the four prompts.

(2) You will be told the current (default) filename, subject number, task code, and replication. If you wish to accept it, type CR; if not, enter a new value. For example, the first prompt is: "If filename RATESFILE not OK, enter new filename (up to 10 characters, no extension):" Type CR to keep RATESFILE, or "MYNAME" to put data in MYNAME (the appropriate extension is added automatically).

3.4. Collecting Subject Ratings

3.4.1. Instructions

Subjects read the rating scale definitions (Appendix A) and the instructions. A copy of the scales is included in Appendix B, for use in briefing subjects. Section 6 contains generic instructions for the keyboard version, and Section 8 contains instructions for the mouse version. Some modifications may be necessary depending on your situation.

3.4.2. Familiarization

Subjects practice using the rating scales after performing a few tasks, to insure that they have developed a standard technique for dealing with the scales.

3.4.3. Using the Program

Subjects perform the experimental task, providing ratings on the six subscales following all task conditions of interest. After the experimenter's initial data input, the six scales will be displayed to the subject on the CRT, along with instructions telling how

to move the pointer, stop its movement, and mark the scale, using either the keyboard arrow keys, or the mouse. Important points to note:

(1) The arrow must be moved before it can be marked;

(2) The point of the arrow indicates the place where the scale is marked. The selected position is indicated by a vertical line.

3.5. Collecting Subject Weights

Subjects complete the "Sources-of-Workload Evaluation" once for each task or group of tasks included in the experiment that share a common structure (although difficulty levels may vary). Subject instructions for doing the Sources-of-Workload Evaluation are in Sections 7 (keyboard version) and 9 (mouse version). After the inital data is input, 15 pairs of scale titles appear, in random order, on the subject's monitor. The subject selects the top or bottom member of each pair by pressing either the correct key on the keyboard or button on the mouse. For the keyboard version, "1" selects the top title and "2" selects the bottom title. A backspace erases a choice, and a CR enters it. For the mouse version, the left button selects the top title and the right button selects the bottom title. A pointer shows which title was selected. Pressing that button again enters the choice. If the subjects changes his or her mind, s/he can press the other button to cancel the first choice, and then press that other button twice to select and enter the new choice. After a choice is entered, a new pair of scale titles is presented.

(Note that the exact time when the weights are obtained is not critical. However, in order for them to provide useful information, they must be obtained after at least some exposure to the relevant task conditions.)

3.6. Summary

Following this procedure, you should end up with:

(1) One or more .RAT files, with at least one rating per experimental condition for each subject, and possibly more.

(2) One or more .WGT files, with a set of workload weights for each subject for each *group* of similar tasks.

4. DATA ANALYSIS PROCEDURE

The COMBINE program is used to produce weighted workload scores. The program takes the specified .RAT file which contains one or more sets of raw ratings for the six subscales, and applies the weights from the requested WGT file to these ratings. It produces a .WWL file which contains subscripts (subject number, condition code, replication), the original sets of ratings, plus a weighted workload score for each of these ratings. The format of this file (for purposes of input to a data analysis program) is as follows: skip the first three lines; then each line has the format 12X, 7I4. The first 12 spaces contain the three identifiers (4 characters each), and the 7 integers are the six raw weights and the weighted workload score.

4.1. Using the Program

Two kinds of information are entered by the experimenter: the relevant filenames, and the experimental variables on which weights vary.

(1) Start the program by typing "Combine," and respond to the prompts.

(2) Enter the name of the ratings file (without extension). (You will get an error message if the file does not exist.)

(3) Enter the name of the weights file to be applied to these ratings. (Again, you will be notified if the file does not exist.)

(4) Enter the name of the new file to be created, which will contain the weighted workload scores (the .WWL extension will be added automatically). (You will be notified if this file already exists, to prevent writing over it.)

(5) You will be asked if weights are conditional on subjects, conditions, and/or replications; in each case, answer "y" or "n." Weights usually *are* conditional on subjects (if you gathered different weights from each subject), and usually are *not* conditional on replications (unless you gathered different weights partway through the experiment). As discussed above, they may or may not be conditional on experimental conditions.

(6) The resulting .WWL file will contain the original ratings and the weighted workload scores (WWL). The ratings can be used as a dependent measure in whatever type of analysis the experimenter chooses.

Figure 1 depicts the composition of a weighted workload score graphically. The bar graph on the left represents six subscale ratings. The width of the subscale bars reflects the importance of each factor (its weight) and the height represents the magnitude of each factor (its rating) in a particular task. The weighted workload score (the bar on the right) represents the average *area* of the subscale bars. If you have any questions, comments, or suggestions about the NASA Task Load Index, please do not hesitate to call us. This procedure is still under evaluation and we are always looking for new ideas.

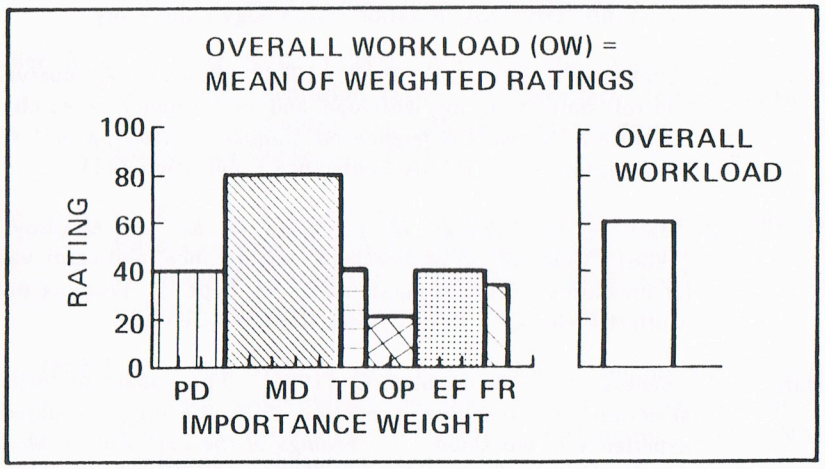


Figure 1: Graphic example of the composition of a weighted workload score.

5. BIBLIOGRAPHY

Biferno, M. A. (1985). *Mental workload measurement: Event-related potentials and ratings of workload and fatigue* (NASA CR 177354). Moffett Field, CA: NASA Ames Research Center.

Bortolussi, M. R., Kantowitz, B. H., & Hart, S. G. (1985). Measuring pilot workload in a motion base trainer: A comparison of four techniques. In R. S. Jensen & J. Adrion (Eds.), *Proceedings of the Third Symposium on Aviation Psychology* (pp. 263-270). Columbus, OH: OSU Aviation Psychology Laboratory.

Hart, S. G., Battiste, V., & Lester, P. T. (1984). Popcorn: A supervisory control simulation for workload and performance research. In *Twentieth Annual Conference on Manual Control* (pp. 431-454). Washington, D.C.: NASA Conference Publication 2341.

Hart, S. G., Battiste, V., Chesney, M. A., Ward, M. M., and McElroy, M. (1986). Comparison of workload, performance, and cardiovascular measures: Type A personalities vs Type B. Working paper. Moffett Field, CA: NASA Ames Research Center.

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* (pp. 732-736). Santa Monica, CA: Human Factors Society.

Hart, S. G., Shively, R. J., Vidulich, M. A., & Miller, R. C. (1986). The effects of stimulus modality and task integrality: Predicting dual-task performance and workload from single-task levels. In *Twenty-First Annual Conference on Manual Control* (pp. 5.1-5.18). Washington, D. C.: NASA Conference Publication 2428.

Hart, S. G., & Staveland, L. E. (In press). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam, The Netherlands: Elsevier.

Haworth, L. A., Bivens, C. C., & Shively, R. J. (1986). An investigation of single-piloted advanced cockpit and control configurations for nap-of-the-earth helicopter combat mission tasks. *Proceedings of the 1986 Meeting of the American Helicopter Society* (pp. 657-672). Washington, D.C.

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 workload. In *Twentieth Annual Conference on Manual Control* (pp. 359-372). Washington, D.C.: NASA Conference Publication 2341.

Miller, R. C., & Hart, S. G. (1984). Assessing the subjective workload of directional orientation tasks. In *Twentieth Annual Conference on Manual Control* (pp. 85-95). Washington, D.C.: NASA Conference Publication 2341.

Mosier, K. L., & Hart, S. G. (1986). Levels of information processing in a Fitts Law task (LIPFitts). In *Twenty-First Annual Conference on Manual Control* (pp. 4.1-4.15). Washington, D.C.: NASA Conference Publication 2428.

Staveland, L. E., Hart, S. G., & Yeh, Y.-Y. (1986). Memory and subjective workload assessment. In *Twenty-First Annual Meeting on Manual Control*. (pp. 7.1-7.13). Washington, D.C.: NASA Conference Publication 2428.

Vidulich, M. A., & Tsang, P. S. (1985a). Techniques of subjective workload assessment: A comparison of two methodologies. In R. S. Jensen & J. Adrion (Eds.), *Proceedings of the Third Symposium on Aviation Psychology* (pp. 239-246). Columbus, OH: OSU Aviation Psychology Laboratory.

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

Vidulich, M. A., & Tsang, P. S. (1986). Collecting NASA Workload Ratings: A Paper and Pencil Package. Working Paper. Moffett Field, CA: NASA Ames Research Center.

Vidulich, M. A. & Tsang, P. S. (in press). Techniques of subjective workload assessment: A comparison of SWAT and the NASA Bipolar Method. *Ergonomics*.

Yeh, Y.-Y., & Wickens, C. D. (1985). The effect of varying task difficulty on subjective workload. In *Proceedings of the Human Factors Society 29th Annual Meeting*. (pp. 765-769). Santa Monica, CA: Human Factors Society.

6. SUBJECT INSTRUCTIONS: RATINGS (Keyboard Version)

We are not only interested in assessing your performance but also the experiences you had during the different task conditions. Right now we are going to describe the technique that will be used to examine your experiences. In the most general sense we are examining the "workload" you experienced. Workload is a difficult concept to define precisely, but a simple one to understand generally. The factors that influence your experience of workload may come from the task itself, your feelings about your own performance, how much effort you put in, or the stress and frustration you felt. The workload contributed by different task elements may change as you get more familiar with a task, perform easier or harder versions of it, or move from one task to another. Physical components of workload are relatively easy to conceptualize and evaluate. However, the mental components of workload may be more difficult to measure.

Since workload is something that is experienced individually by each person, there are no effective "rulers" that can be used to estimate the workload of different activities. One way to find out about workload is to ask people to describe the feelings they experienced. Because workload may be caused by many different factors, we would like you to evaluate several of them individually rather than lumping them into a single global evaluation of overall workload. This set of six rating scales was developed for you to use in evaluating your experiences during different tasks. Please read the descriptions of the scales carefully. If you have a question about any of the scales in the table, please ask me about it. It is extremely important that they be clear to you. You may keep the descriptions with you for reference during the experiment.

After performing each task, six rating scales will be displayed. You will evaluate the task by marking each scale at the point which matches your experience. Each line has two endpoint descriptors that describe the scale. Note that "own performance" goes from "good" on the left to "bad" on the right. This order has been confusing for some people. Move the arrow with the right and left arrow keys until it points at the desired location. Stop it by pressing the up arrow key. Press the down arrow key to enter your selection. Please consider your responses carefully in distinguishing among the task conditions. Consider each scale individually. Your ratings will play an important role in the evaluation being conducted, thus, your active participation is essential to the success of this experiment, and is greatly appreciated.

7. SUBJECT INSTRUCTIONS: RATINGS (Mouse Version)

We are not only interested in assessing your performance but also the experiences you had during the different task conditions. Right now we are going to describe the technique that will be used to examine your experiences. In the most general sense we are examining the "workload" you experienced. Workload is a difficult concept to define precisely, but a simple one to understand generally. The factors that influence your experience of workload may come from the task itself, your feelings about your own performance, how much effort you put in, or the stress and frustration you felt. The workload contributed by different task elements may change as you get more familiar with a task, perform easier or harder versions of it, or move from one task to another. Physical components of workload are relatively easy to conceptualize and evaluate. However, the mental components of workload may be more difficult to measure.

Since workload is something that is experienced individually by each person, there are no effective "rulers" that can be used to estimate the workload of different activities. One way to find out about workload is to ask people to describe the feelings they experienced. Because workload may be caused by many different factors, we would like you to evaluate several of them individually rather than lumping them into a single global evaluation of overall workload. This set of six rating scales was developed for you to use in evaluating your experiences during different tasks. Please read the descriptions of the scales carefully. If you have a question about any of the scales in the table, please ask me about it. It is extremely important that they be clear to you. You may keep the descriptions with you for reference during the experiment.

After performing the task, six rating scales will be displayed. You will evaluate the task by marking each scale at the point which matches your experience. Each line has two endpoint descriptors that describe the scale. Note that "own performance" goes from "good" on the left to "bad" on the right. This order has been confusing for some people. Move the arrow to the right or left with the mouse until it points at the desired location. When you are satisfied, press either button to enter your selection. Please consider your responses carefully in distinguishing among the task conditions. Consider each scale individually. Your ratings will play an important role in the evaluation being conducted, thus, your active participation is essential to the success of this experiment, and is greatly appreciated.

8. SUBJECT INSTRUCTIONS: SOURCES-OF-WORKLOAD EVALUATION (Keyboard Version)

Throughout this experiment the rating scales are used to assess your experiences in the different task conditions. Scales of this sort are extremely useful, but their utility suffers from the tendency people have to interpret them in individual ways. For example, some people feel that mental or temporal demands are the essential aspects of workload regardless of the effort they expended or the performance they achieved. Others feel that if they performed well the workload must have been low, and vice versa. Yet others feel that effort or feelings of frustration are the most important factors in workload; and so on. The results of previous studies have found every conceivable pattern of values. In addition, the factors that create levels of workload differ depending on the task. For example, some tasks might be difficult because they must be completed very quickly. Others may seem easy or hard because of the intensity of mental or physical effort required. Yet others feel difficult because they cannot be performed well, no matter how much effort is expended.

The evaluation you are about to perform is a technique that has been developed by NASA to assess the relative importance of six factors in determining how much workload you experienced. The procedure is simple: You will be presented with a series of pairs of rating scale titles (for example, Effort vs. Mental Demands) and asked to choose which of the items was more important to your experience of workload in the task(s) that you just performed. Each pair of scale titles will appear separately on the screen. Select the Scale Title that represents the more important contributor to workload for the specific task(s) you performed in this experiment.

Press "1" to select the top item in the pair, and "2" to select the bottom item. If you change your mind, press backspace to erase your choice. Press carriage return to enter it. After the carriage return, a new pair of scale titles will appear.

After you have finished the entire series we will be able to use the pattern of your choices to create a weighted combination of the ratings from that task into a summary workload score. Please consider your choices carefully and make them consistent with how you used the rating scales during the particular task you were asked to evaluate. Don't think that there is any *correct* pattern; we are only interested in your opinions. If you have any questions, please ask them now. Thank you for your participation.

9. SUBJECT INSTRUCTIONS: SOURCES-OF-WORKLOAD EVALUATION (Mouse Version)

Throughout this experiment the rating scales are used to assess your experiences in the different task conditions. Scales of this sort are extremely useful, but their utility suffers from the tendency people have to interpret them in individual ways. For example, some people feel that mental or temporal demands are the essential aspects of workload regardless of the effort they expended or the performance they achieved. Others feel that if they performed well the workload must have been low, and vice versa. Yet others feel that effort or feelings of frustration are the most important factors in workload; and so on. The results of previous studies have already found every conceivable pattern of values. In addition, the factors that create levels of workload differ depending on the task. For example, some tasks might be difficult because they must be completed very quickly. Others may seem easy or hard because of the intensity of mental or physical effort required. Yet others feel difficult because they cannot be performed well, no matter how much effort is expended.

The evaluation you are about to perform is a technique that has been developed by NASA to assess the relative importance of six factors in determining how much workload you experienced. The procedure is simple: You will be presented with a series of pairs of rating scale titles (for example, Effort vs. Mental Demands) and asked to choose which of the items was more important to your experience of workload in the task(s) that you just performed. Each pair of scale titles will appear separately on the screen. Select the Scale Title that represents the more important contributor to workload for the specific task(s) you performed in this experiment.

Press the left button to select the top item in the pair, and the right button to select the bottom item. A pointer shows which title was selected. To enter that choice, press the button again, and a new pair of titles will appear. If you change your mind, press the other button to cancel your first cohice, and then start over.

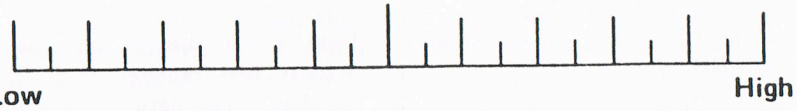
After you have finished the entire series we will be able to use the pattern of your choices to create a weighted combination of the ratings from that task into a summary workload score. Please consider your choices carefully and make them consistent with how you used the rating scales during the particular task you were asked to evaluate. Don't think that there is any *correct* pattern; we are only interested in your opinions. If you have any questions, please ask them now. Thank you for your participation.

RATING SCALE DEFINITIONS

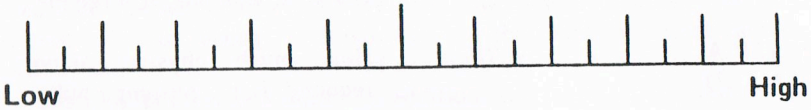
Title	Endpoints	Descriptions
MENTAL DEMAND	*Low/High*	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	*Low/High*	How much physical 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?
EFFORT	*Low/High*	How hard did you have to work (mentally and physically) to accomplish your level of performance?
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?
FRUSTRATION LEVEL	*Low/High*	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Appendix B.

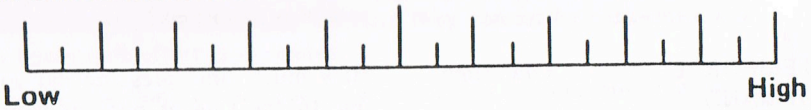
MENTAL DEMAND



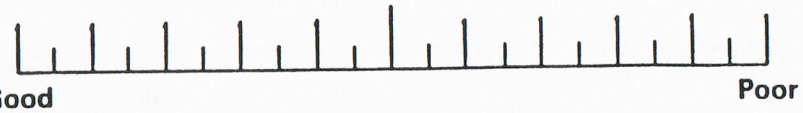
PHYSICAL DEMAND



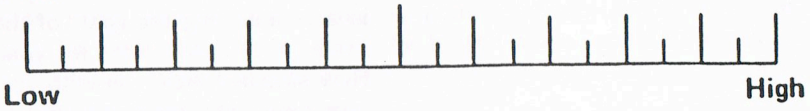
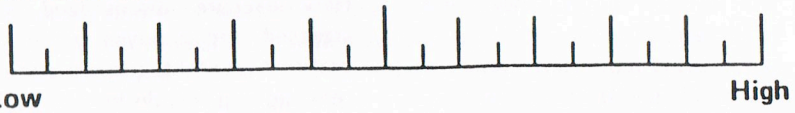
TEMPORAL DEMAND



PERFORMANCE



EFFORT



FRUSTRATION