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

NASA Task Load Index (NASA-TLX)
Version 1.0

Paper and Pencil Package

This booklet contains the materials necessary to collect subjective workload assessments with the NASA Task Load Index. 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 booklet 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. 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 (Apppendix B). Each pair is presented on a card. Subjects circle the member of each pair that contributed more to the workload of that task. The number of times that each factor is selected is tallied. 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 scales are presented on a rating sheet (Appendix C). Subjects respond by marking each scale at the desired location. In operational situations, rating sheets or verbal responses are more practical, while a computerized version (available from NASA Ames Research Center) is more efficient for most simulation and laboratory settings. Ratings may be obtained either during a task, after task segments, or following an entire task. Each scale is presented as a 12-cm line divided into 20 equal intervals anchored by bipolar descriptors (e.g., High/Low). The 21 vertical tick marks on each scale divide the scale from 0 to 100 in increments of 5. If a subject marks between two ticks, the value of the right tick is used (i.e., round up).

2.4. Weighting and Averaging Procedure

The overall workload score for each subject is computed 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). (See Appendix D and E for a sample Tally Sheet and Worksheet.)

3. EXPERIMENTAL PROCEDURE

The usual sequence of events for collecting data with the NASA Task Load Index is as follows:

3.1. Instructions

Subjects read the scale definitions and instructions. A set of generic instructions is included in Section 6. Some modifications may be necessary depending on your situation.

3.2. 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.

3.3. Ratings

Subjects perform the experimental tasks, providing ratings on the six subscales following all task conditions of interest. The number of rating sheets needed equals the number of subjects X

the number of task conditions (including practice).

3.4. 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). For example, in an experiment with several memory tasks and several tracking tasks, two Sources-of-Workload Evaluations would be performed: one for the memory tasks and one for the tracking tasks. One set of cards should be made in advance of the experiment for each subject X evaluation condition combination. The pairs of factors should be cut apart and presented individually in a different, randomly selected, order to each subject. Subject instructions for doing the Sources-of-Workload Evaluation are in Section 7. (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.)

3.5. Summary

Following this procedure, you should end up with: (1) a set of workload weights from each subject for each group of similar tasks, and (2) at least one rating sheet for each subject for each experimental task. Typically, we have run within-subject experiments and therefore ended up with a larger number of rating sheets for each subject.

To conserve paper and speed up the subsequent analysis, we often enclose the Rating Sheet and the Sources-of-Workload comparison cards in clear plastic. Subjects mark the scales with an erasable felt tip marker. Immediately after they are marked, the experimenter transfers the responses onto the appropriate form or worksheet. Then the plastic sheets are cleaned and reused. If this procedure is followed, *DOUBLE CHECK YOURSELF BEFORE ERASING THE SUBJECT'S RESPONSES!*

4. DATA ANALYSIS PROCEDURE

The procedure for computing a weighted workload score follows:

1.1. Tally Sheet

For each subject, the "Sources-of-Workload Tally Sheet" (Appendix D) is used to compute the weight for each factor. The scorer simply leafs through the evaluation cards and puts a mark on the appropriate row of the tally column for each response of the subject (e.g., each time the subject circled "Mental Demand" on a comparison card, the experimenter would put a mark in the "Mental Demand" row of the tally column). After going through the Sources-of-Workload evaluation, the experimenter adds the tallies for each scale and writes the totals in the "Weight" column.

1.2. Worksheet

The Weight column from the tally sheet is then transferred to the "Weighted Rating Worksheet" (Appendix E). Each subject would have his or her individual workload parameters count placed on a separate worksheet for the appropriate task or set of similar tasks. If subjects rated more than one task, the appropriate number of copies of the worksheet should be made. Ratings are placed in the "Raw Rating" column of the worksheet. The "Adjusted Rating" is formed by multiplying the Raw Rating by the Sources-of-Workload Weight. The adjusted ratings are summed across the different scales. The sum is divided by 15 to obtain the overall weighted workload score for the subject in that one task condition.

The weighted ratings are then used as a dependent measure in whatever type of analyses 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.

1.3. Summary

The above procedure, although simple, can be laborious for a large experiment. Thus it is highly advantageous to computerize the procedure. A set of programs that run on IBM-PC compatible machines has been

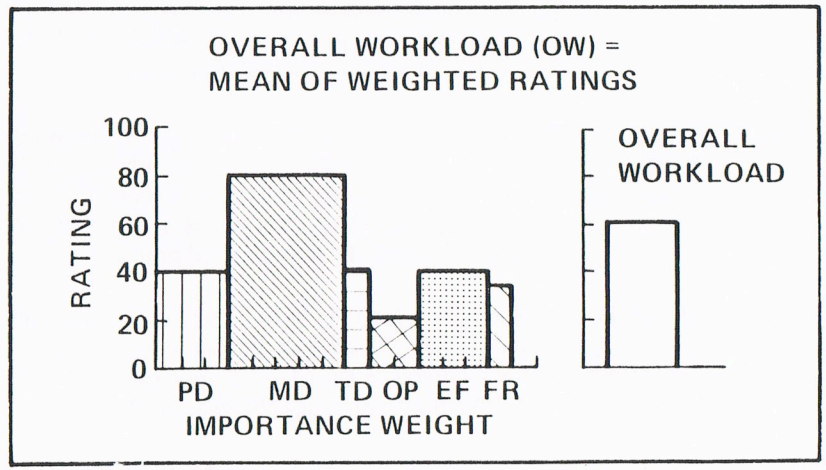


Figure 1: Graphic example of the composition of a weighted workload score

written to gather ratings and weights, and compute the weighted workload scores. These are available upon request from NASA Ames Research Center. However, if this is not a viable option, all the necessary materials are included in this booklet. If you have any questions, comments, or suggestions please do not hesitate to contact us. This procedure is still under evaluation and we are always looking for new ideas.

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: RATING SCALES

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 of the tasks, you will be given a sheet of rating scales. You will evaluate the task by putting an "X" on each of the six scales 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. Please consider your responses carefully in distinguishing among the different 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 by all of us.

7. SUBJECT INSTRUCTIONS: SOURCES-OF-WORKLOAD EVALUATION

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 on a given task or the level of performance they achieved. Others feel that if they performed well the workload must have been low and if they performed badly it must have been high. 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 on a separate card.

Circle the Scale Title that represents the more important contributor to workload for the specific task(s) you performed in this experiment.

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. Otherwise, start whenever you are ready. 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?
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?

Appendix B.

Sources-of-Workload Comparison Cards

Effort

or

Performance

Temporal Demand

or

Frustration

Temporal Demand

or

Effort

Physical Demand

or

Frustration

Performance

or

Frustration

Physical Demand

or

Temporal Demand

Physical Demand

or

Performance

Temporal Demand

or

Mental Demand

Frustration

or

Effort

Performance

or

Mental Demand

Performance

or

Temporal Demand

Mental Demand

or

Effort

Mental Demand

or

Physical Demand

Effort

or

Physical Demand

Frustration

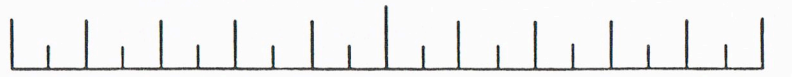
or

Mental Demand

Subject ID: _____ Task ID: _____

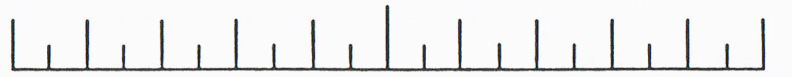
RATING SHEET

MENTAL DEMAND



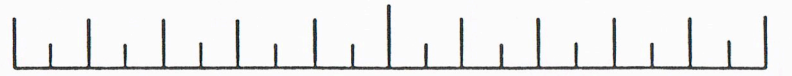
Low High

PHYSICAL DEMAND



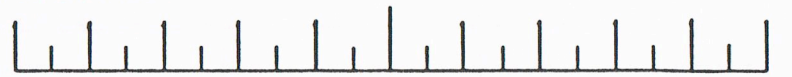
Low High

TEMPORAL DEMAND



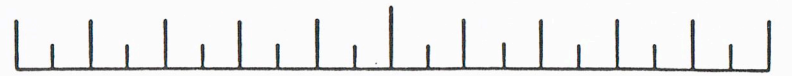
Low High

PERFORMANCE



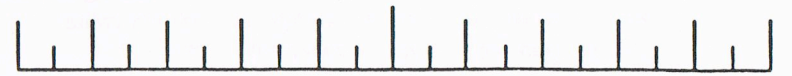
Good Poor

EFFORT



Low High

FRUSTRATION



Low High

Appendix D.

Subject ID: _____ **Date:** _____

SOURCES-OF-WORKLOAD TALLY SHEET		
Scale Title	*Tally*	*Weight*
MENTAL DEMAND		
PHYSICAL DEMAND		
TEMPORAL DEMAND		
PERFORMANCE		
EFFORT		
FRUSTRATION		

Total count = _____

(NOTE - The total count is included as a check. If the total count is not equal to 15, then something has been miscounted. Also, no weight can have a value greater than 5.)

Appendix E.

Subject ID: _____ **Task ID:** _____

WEIGHTED RATING WORKSHEET			
Scale Title	*Weight*	*Raw Rating*	*Adjusted Rating* (Weight X Raw)
MENTAL DEMAND			
PHYSICAL DEMAND			
TEMPORAL DEMAND			
PERFORMANCE			
EFFORT			
FRUSTRATION			

Sum of "Adjusted Rating" Column = _____

WEIGHTED RATING =
[i.e., (Sum of Adjusted Ratings)/15]