NASA-TASK LOAD INDEX (NASA-TLX); 20 YEARS LATER

Sandra G. Hart
NASA-Ames Research Center
Moffett Field, CA

ABSTRACT

NASA-TLX is a multi-dimensional scale designed to obtain workload estimates from one or more operators while they are performing a task or immediately afterwards. The years of research that preceded subscale selection and the weighted averaging approach resulted in a tool that has proven to be reasonably easy to use and reliably sensitive to experimentally important manipulations over the past 20 years. Its use has spread far beyond its original application (aviation), focus (crew complement), and language (English). This survey of 550 studies in which NASA-TLX was used or reviewed was undertaken to provide a resource for a new generation of users. The goal was to summarize the environments in which it has been applied, the types of activities the raters performed, other variables that were measured that did (or did not) covary, methodological issues, and lessons learned

BACKGROUND

Workload is a term that represents the cost of accomplishing mission requirements for the human operator. If people could accomplish everything they are expected to do quickly, accurately, and reliably using available resources, the concept would have little practical importance. Since they often cannot, or the human cost (*e.g.*, fatigue, stress, illness, and accidents) of maintaining performance is unacceptably high, designers, manufacturers, managers, and operators, who are ultimately interested in system performance, need answers about operator workload at all stages of system design and operation. The many definitions that exist in the psychological literature are a testament to the complexity of the construct as are the growing number of causes, consequences and symptoms that have been identified. Given the confusion among the "experts", it seems equally likely that people who are asked to provide ratings will have a similar range of opinions and apply the same label (workload) to very different aspects of their experiences.

For this reason, the NASA Task Load Index (NASA-TLX) consists of six subscales that represent somewhat independent clusters of variables: Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance. (Appendix). The assumption is that some combination of these dimensions are likely to represent the "workload" experienced by most people performing most tasks. These dimensions were selected after an extensive analysis of the primary factors that do (and do not) define the subjective experience of workload for different people performing a variety of activities ranging from simple laboratory tasks to flying an aircraft. Coincidentally, these dimensions also correspond to various theories that equate workload with the magnitude of the demands imposed on the operator, physical, mental, and emotional responses to those demands or the operator's ability to meet those demands.

A weighting scheme was introduced to take such individual differences into account when computing an overall workload score (Figure 1). Essentially, overall workload represents the total areas of the six bars. The weights are derived for each participant at the beginning of a study by requiring simple decisions about which member of each paired combination of the 6 dimensions are more related to their personal definition of workload. Each subscale rating provided by that person during the study is then multiplied by the appropriate weight, developing a composite tailored to individual workload definitions. The benefit of this weighting scheme was an increase in sensitivity (to relevant variables) and a decrease in between-rater variability. The development and theoretical rationale for the scale were described in a chapter published in 1988 by Hart & Staveland.

Since its introduction, NASA-TLX has been translated into more than a dozen languages, administered verbally, in writing, or by computer, and modified in a variety of ways. It has also been subjected to a number of independent evaluations in which its reliability, sensitivity, and utility were

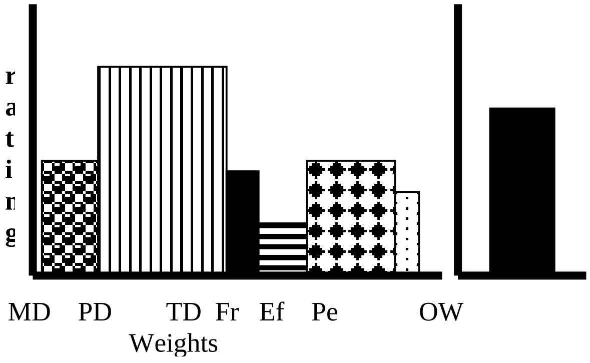


Fig 1: Graphic Representation of weighted subscale ratings and an overall workload value

assessed and compared to other methods of measuring workload.

SELECTING STUDIES

A number of resources were used to identify the articles to be reviewed. Simply "Googling" the phrase "NASA TLX" returned 82,900 citations, 44,000 of which were in English. Entering the phrase "NASA-TLX", returned 25,000 and 19,800 citations, respectively. A spot check revealed that most were legitimate, but that many were duplicates. Since wading through all of them would be an impossible task, I turned to Google Scholar which offered a more manageable return: 1200 for "NASA-TLX" and 1670 for "NASA TLX". There were only 446 returns for the seminal reference, "Hart, S. G. & Staveland, L. E." possibly because the book has been out of print for years and an increasing number of users are simply citing the measure without reference, much as they would cite the use of any other measure (e.g., reaction time or percent correct).

Rather than focusing solely on refereed journal articles, I decided to draw from a variety of sources to get as broad a sample of NASA-TLX users as possible. Included are more than 15 years of journal articles and conference proceedings from the Human Factors and Ergonomics Society, government reports from a number of countries, book chapters, conference proceedings, and papers published in a variety of other journals. I also searched the internet for a new generation of users who publish online or are from non-English-speaking countries. I was able to review 550 articles in the time available - - a reasonable cross-section of what has been done but not a statistically random sample.

CODING SCHEME

I categorized each article with respect to the country in which the first author lived, the nature of the organization(s) that performed and funded the work, the domain to which it applied (e.g., aviation, automobiles, medicine, combat, etc), its focus (e.g., interface design, automation, training, model validation), and covariates (e.g., situation awareness, experience, fatigue).

Region	% papers
Africa	0
Asia	4
Australia/NZ	2
Europe	16
Middle East	1
No America	77
So America	1

Table 1: Region of world in which first author lived

Finally, the circumstances in which NASA-TLX (and its subscales) were found to be appropriately sensitive to experimentally manipulated variables were summarized and compared to the relative successes (and failures) of the variations that have been developed and used over the years. This summary is too lengthy to include in this brief paper but will be available on the website.

RESULTS

Language

Table 1 summarizes the region of the world in which the first authors lived (note: this and other information was not always available, especially for articles found on the internet so I summarized what was available). As might be expected for a scale originally written in English, the original English version was used in most of the studies. However, as the use of NASA-TLX spread, it was translated into other languages, necessitating re-validation to demonstrate its validity and sensitivity in other cultures. In 28 of the articles, the author stated that the scale had been translated into some other language and was being evaluated or applied. I suspect that was the case in additional studies, but could not verify it. The initial translations were into French (*cf*, Rubio, Diaz, Martin, Puente, 2004), German (*cf*, Sepehr, 1988) and Japanese (*cf*, Haga, Shinoda, & Kokubun, 2002). Recent translations include Korean, Spanish, Portugese, Norwegian, and Chinese.

From the available information, it appears that the scale titles, anchors, and definitions were not translated literally, but rather in a manner appropriate for each culture and language. My own language skills limit my ability to comment on the quality of the translations, but it appears that most of the foreign-language versions have been deemed a success and are being used.

	Performing Org		
Funding Org	**Govt**	**Industry**	**Univ**
US Air Force	24	2	12
US Army	14	9	27
US FAA	16	4	5
US NASA	34	4	27
US Navy	2	6	2
US govt other	5	4	23
British govt	4		3
Canadian govt	12		6
European Union	17		3
Euro Sp Agenc	1		3
French govt	2		
German govt	4		3
Norwegian govt	2		
Swedish govt	3		3
Other non-US	3		2
Unknown		38	178

Table 2: Organizations that performed and/or funded the studies

Performing/Funding Organizations

Most of the studies were performed by US (17%) or other government research labs (9%) or by universities alone (32%) or in collaboration with and/or funded by agencies of the US (17%) or other (12%) governments. Only 12% of the studies that I reviewed were performed by industry with internal or government funding. (Table 2)

Focus

Most of the studies addressed some sort of question about interface design or evaluation: Visual and/or auditory displays (31%), vocal and/or manual input devices (11%), virtual or augmented vision (6%). In addition, these and other studies also examined the impact of underlying systems such as automation and decision aids (26%), digital data link (3%), Caution, Advisory and Warning systems (4%), and new types of information on operator workload.

In the same or different studies, the relationship between NASA-TLX ratings and other factors also relevant to successful performance were assessed: Teamwork (6%), crew size (1%), fatigue (2%), stress (3%), trust (2%), age (1%), personality (2%), experience (4%), and disability/illness (1%). The most popular covariate was Situation Awareness (SA), cited in 7% of the studies. SA is as ill-defined a construct and as descriptively and practically useful as is workload itself. However, the correlations between SA and workload found in different studies were positive, negative, or none (*cf*, Hansman). In fact, it was suggested that SA is simply a consequence of workload and not an independent phenomenon (Hendy, 1995).

Most studies included measures of performance and many also included measures of physiological (e.g., cardiovascular, muscular, skin, brain) function thought to index different aspects of workload were included in 6% of the studies (*cf*, Prinzel, Pope, and Freeman, 2001). NASA-TLX ratings may or may not covary with measures of performance (dissociation). For example, the cost of performing well in a difficult task may be an unacceptably high level of workload. On the other hand, the workload cost of performing an apparently undemanding vigilance task can be extremely high prompted by boredom (Warm, Dember, Hancock, 1996) unless ameliorated by improved design. It is for just this reason that designers need information about workload as well as performance.

Domain

The majority of the studies targeted a specific operational environment, even though the actual study was performed in a laboratory (10%) or simulation environment. Since NASA-TLX was initially designed for use in aviation, it is not surprising that many of the studies were focused on Air Traffic Control (10%) or civilian (12%) and military (5%) cockpits. Its use spread next to the military for armored vehicles (2%), soldiers (3%), and command and control (3%) and then into power plants (3%), various forms of remote control including robotics, unmanned vehicles, and teleoperation (5%), and space applications (2%). In the last ten years, an increasing number of studies have focused on automobile drivers (8%), the medical profession (4%), and users of computers (7%) or personal, portable technologies such as cell phones (4%).

Activity

Within and across studies, there are common sorts of human activities that are of interest in assessing workload. They include manual control tasks such as flying (14%), driving (9%), and data entry (10%), visual and auditory monitoring (3%), decision making (3%), teamwork (6%) , communications (2%) and so on. It is because both laboratory and real-world tasks have these basic human activities in common that laboratory research results and well-designed measurement tools can be applied a across domains.

Methodological Issues

Nearly 25% of the articles described efforts to develop, evaluate, or compare new and/or existing subjective, performance-based, and physiological measures of workload for a specific type of application or country. In many cases, the reviews or websites simply packaged existing knowledge, making it more readily available (Federal Aviation Administration, 2006). Other reviews offered recommendations for a specific occupation (*e.g.*, powerplant operation; *cf*, Lang *et al*, 2002), activity (e.g., mobile air defense; *cf*, Bittner, *et al*, 1989), user community (e.g., cell phone users; *cf*, Cockburn & Siresena, 2003) or locale (e.g., air traffic management in Europe; *cf*, Straeter & Barbarino, 2004).

In other articles, the authors propose and apply a modified version of the original scale. Some add subscales (6 articles), while others delete them (12 articles) or redefine the existing subscales to improve the relevance to the target task or experimental questions. While increasing the fit between the generic NASA-TLX labels and definitions to a situation can be an excellent strategy, it does require establishing the validity, sensitivity, and reliability of the new instrument before using it. A good example of such an effort may be found in Park & Cha (1998) where several variants of NASA-TLX were evaluated for use by Korean drivers. A practical problem with adding, deleting, and re-defining subscales and continuing to refer to the result as "NASA-TLX" even though the new scale shares only a passing similarity with the original is that it makes if difficult to summarize the circumstances under which the original scale is and is not useful.

The most common modification made to NASA-TLX has been to eliminate the weighting process all together or weighting the subscales and then analyzing them individually. The former has been referred to as Raw TLX (RTLX) and has gained some popularity because it is simpler to apply; the ratings are simply averaged or added to create an estimate of overall workload. In the 29 studies in which RTLX was compared to the original version, it was found to be either more sensitive (Hendy, Hamilton, & Landry, 1993), less sensitive (Liu & Wickens, 1994), or equally sensitive (Byers, Bittner, Hill, 1989), so it seems you can take your pick.

The other common variation is to analyze subscale ratings *instead of* generating a single overall workload score. This was done in at least 40 of the studies I reviewed. In addition, individual subscale analyses were performed *in addition* to assessing overall workload in nearly 20% of the studies. Both of these approaches demonstrate one of the continuing strengths of the scale; the diagnostic value of the component subscales. The component ratings can help designers pinpoint the source of a workload or performance problem.

Other methodological issues of note include the problem of context effect (i.e., TLX ratings of one task may be influenced by significantly different experiences immediately before), range or anchor effects (raters do not use the whole range of the scale and/or tailor their use of the scale to the set of experiences they have in the experimental environment). All of these problems are typical of subjective ratings in general, and can be avoided by providing explicit experiences or instructions to serve as anchors and being careful to control context effects. Finally, users continue to point out that the subscales are often significantly correlated with each other. I believe that this simply illustrates the fact they are all measuring some aspect of the same underlying entity.

The final methodological issue has to do with the elusive workload "redline"; a point on the scale that indicates when workload is not only high, but *too* high. Redlines have been proposed for at least one other scale, but few studies have proposed one for NASA-TLX (*cf*, Hoffman, Pene, & Rognin, Zeghal, 2003). Given the relative nature of subjective ratings, I still feel we are a long way from defining a useful "redline" that can be applied across applications and tasks. Perhaps an inclusive meta-analysis of the hundreds of studies that been done might provide the information needed.

SUMMARY

In the past year, the software for administering the NASA-TLX received a long-overdue modernization and a website has been established from which the software, articles, instructions, and this survey can be downloaded. The goal was to make the wealth of lessons learned from previous users of NASA-TLX readily available. Not only is it important to know which questions it has been successful in answering but also the situations in which it has failed. These analyses of success/failure have been necessarily simplistic, as it is almost impossible to distinguish instances in which NASA-TLX was not sensitive to an experimental manipulation (that really did influence workload) from one in which no significant rating differences reflected reality. It is hoped that the accumulated evidence from many different applications will provide insight.

After nearly 20 years of use, NASA-TLX has achieved a certain venerability; it is being used as a benchmark against which the efficacy of other measures, theories, or models are judged. It is described in college text books, taught in university courses, and recommended for use in situations as diverse as aircraft certification, operating rooms, nuclear power plant control rooms, simulated combat, and website design. On the other hand, the continuing series of evaluations, modifications, extensions, and applications to new situations seem to be keeping it young.

REFERENCES

Bittner, A. V., Byers, J. C., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1989) Generic workload ratings of a mobile air defense system (LOS-FH). *Proceedings of the 33rd Annual Meeting of the Human Factors and Ergonomics Society*, 1476-1480. Santa Monica, CA: HFES.

Byers, J. C., Bittner, A. C., & Hill, S. G. (1989) Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? Advances in Industrial Ergonomics and *Safety*. A. Mital (Ed.) Taylor & Francis., 481-485

Cockburn, A., & Siresena, A. (2003) Evaluating mobile text entry with the FASTAPTM keypad. *People and Computers XVII (Vol 2): Conference on Human Computer Interaction*.Bath,: British Computing Soc.

Federal Aviation Admin. (2006)Workbench tools http://www.hf.faa.gov/WorkbenchTools/

Haga, S., Shinoda, H/. Kokubun, M., (2002) Effects of Task Difficulty and Time-on-Task on Mental Workload, *Japanese Psychological Research, 44* (3), 134-143

Hansman, J. (2004) Workload and situation awareness. Cambridge, MA; MIT.

Hart, S. G. & Staveland, L. E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) *Human Mental Workload*. Amsterdam: North Holland Press.

Hendy, K. C. (1995) Situation awareness and workload: Birds of a feather*? Situation Awareness: Limitations and Enhancement in the Aviation Environment*. Neuilly sur Seinne, FR: AGARD 21-1 – 22-7

Hendy, K. C., Hamilton, K. M., & Landry, L. M. (1993) Measuring subjective workload: When is one scale better than many? *Human Factors, 35* (4), 579-601.

Hoffman, E., Pene, N., Rognin, L., Zeghal, K. (2003) Introducing a new spacing instruction, impact of spacing tolerance on flight crew activity. *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*, 174-178, Santa Monica, CA: HFES

Lang, A,. W. Roth, E. M. Bladh, K. Hine, R., (2002) Using a benchmark-referenced approach for validating a powerplant control room: Results from the baseline study, *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society*, 1878-1882, Santa Monica, CA: HFES.

Liu, Y. L., & Wickens, C. D. (1994) Mental workload and cognitive task automaticity - An evaluation of subjective and time-estimation metrics, *Ergonomics, 37* (11), 1843-1854

Park, P., Cha, D. W. (1998) Comparison of subjective mental workload assessment techniques for the evaluation of in-vehicle navigation system utility. Suwon, Korea: Ajou University,

Prinzel, L. J. III Pope, A. T. Freeman, F. G. (2001) *Application of physiological self-regulation and adaptive task allocation techniques for controlling operator hazardous states of awareness*. (NASA TM-2001-211015) Hampton, VA: National Aeronautics and Space Admin.

Rubio, S., Diaz, E., Martin, J., &Puente, J. M. (2004) Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and Workload Profile, *Applied Psychology: An International Review, 53*(1), 61-86.

Straeter, O. & Barbarino (2004) *A tool for the assessment of the impact of change in automated ATM systems on mental workload.* European Organization for the Safety of Air Navigation Report (HRS/ HSP-0050REP-03). Brussels: Eurocontrol.

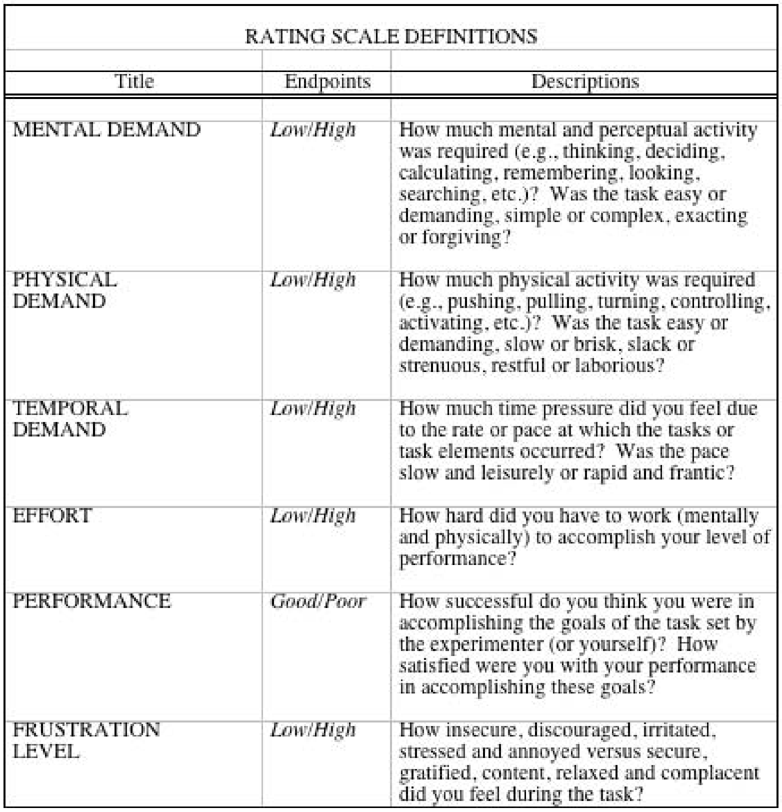
Sepehr, M. M. (1988) Assessment of Subjective Mental Workload Using NASA-Task Load Index. *Proceedings of the 7th European Annual Conference on Human Decision Making and Manual Control*, 69-75.

Warm, J. S. Dember, W. N. Hancock, P. A. (1996) Vigilance and workload in automated systems. *In Automation and Human Performance*. R. Parasuraman and M. Mouloua (Eds) Mahwah, NJ: Erlbaum, 183-200

Xiao Y. M, Wang Z. M, Wang M. Z, & Lan Y. J, The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index (English translation). Chengdu, China: Sichuan U, Institute of Occupational Health

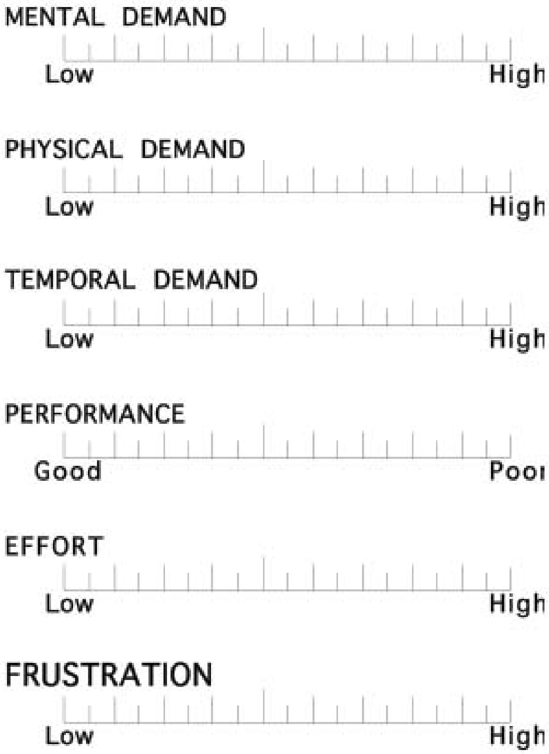
APPENDIX:

Rating Scale and Definition



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?


MENTAL DEMAND
Low
High
PHYSICAL DEMAND
Low
High
TEMPORAL DEMAND
Low
High
PERFORMANCE
Good
Pooı
EFFORT
Low
High
FRUSTRATION
Low
High