HOW HIGH IS HIGH? A META-ANALYSIS OF NASA-TLX GLOBAL WORKLOAD SCORES

Rebecca A. Grier
Institute for Defense Analyses

This paper presents a descriptive analysis of over 1000 global NASA Task Load Index (TLX; Hart & Staveland, 1988) scores from over 200 publications. This analysis is similar to that which was suggested by Hart (2006). The frequency distributions and measures of central tendency presented will aid practitioners in understanding global NASA-TLX scores observed in system tests.

Over 25 years ago, Hart and Staveland (1988) published the NASA Task Load Index (TLX), a multi-dimensional survey based measure of workload. It is now the most cited survey based workload measure. Hart's (2006) retrospective of the first 20 years of the NASA-TLX documented its vast use. Moreover, Google Scholar™ (www.scholar.google.com) reports that there are more than twice as many search results for the NASA-TLX than other survey based workload measures (see Table 1).

Table 1

Number of Approximate Results on Google Scholar™ for Common Workload Measures as of June 2015

Search Term	Results
"NASA TLX"	10,300
SWAT workload	3,620
"Cooper Harper" workload	2,260
"Subjective Workload Dominance"	229
"Bedford workload"	177

As initially published, the NASA-TLX global workload score is the weighted mean of six subscales, mental demand, physical demand, temporal demand, frustration, effort, and performance. That is, participants rate six sub-scales from 0 to 100 based on the experience of each in the preceding task. The weights are obtained by asking the participants to complete a series of 15 paired comparisons (i.e., all combinations of the six sub-scales) by identifying which sub-scale contributed more to their experience of workload in the task. The number of selections (zero to five) is the weight assigned to each sub scale. The resulting weighted mean is the global workload score and can be between 0 (low) and 100 (high).

Hart and Staveland (1988) also documented an extensive research program assessing the reliability and validity of the NASA-TLX. The results showed that the NASA-TLX is more pragmatic and less sensitive to individual differences, than the Subjective Workload Assessment Technique (SWAT). This research also demonstrated that the NASA-TLX is more sensitive to workload differences than the Overall Workload survey (OW; Hart & Staveland, 1988). Other researchers have independently demonstrated the reliability and validity of the NASA-TLX in comparison to other workload measures (e.g., Byers, Bittner, Hill, Zaklad, & Christ, 1988; Bittner, Byers, Hill, Zaklad, & Christ, 1989). Furthermore, other researchers validated its administration via computer (e.g., Sharek, 2011) and verbally (Grant, 2008).

A common modification of the NASA-TLX is referred to as the RTLX or Raw TLX, which is obtained when the sub-scales are averaged without completing the paired comparisons (Hart, 2006). Many studies (e.g., Nygren, 1991; Moroney, Warm, & Dember, 1995) have shown there is a high correlation between the weighted workload score (WWL) and the RTLX.

Hart (2006) noted that a limitation of the NASA-TLX is the interpretation of scores. She further noted that the analysis of the vast amounts of data published could reduce this limitation (Hart, 2006). To that end, the present study defines the range and cumulative frequencies of published scores, which will enable practitioners to state the percentage of scores that have been reported as higher or lower than the observed score.

METHOD

Data Gathering

Due to time limitations, the analyses were limited to the articles found in the publications of the Human Factors and Ergonomics Society (HFES; i.e., Human Factors, Journal of Cognitive Engineering and Decision Making, Proceedings of the Annual Meeting of the Human Factors & Ergonomics Society) and the "Complete TLX Publication List." The latter is a database of publications citing the NASA-TLX updated until approximately 2003 and retrieved from http://humansystems.arc.nasa.gov/groups/tlx/tlxpublications.html.

The "Complete TLX Publication List" (n = 297) contained several inaccurate citations and duplicate entries. After cleaning the list, there were 277 usable publications. Of these, 122 were also HFES publications. Of the remaining 155 unique publications, only 58 were available online in English. These 58 publications were added to the 722 publications from HFES and served as the initial list of 780 publications reviewed.

Of the 780 publications considered, 237 were included in the analysis. Several publications were excluded because they cited the NASA-TLX, but did not report research with it.

Not subject to U.S. copyright restrictions. DOI 10.1177/1541931215591373

Articles that did utilize the NASA-TLX in research were excluded for the following reasons: (1) not reporting the global workload score numerically or in a readable graph, (2) being a duplicate data publication (i.e., data published in a brief format in a conference proceedings and in an expanded format in a journal article), or (3) indicating the NASA-TLX was modified in a way that has not been validated. The goal was to be as inclusive as possible with the publications. However, it was deemed that modifications worthy of comment in the publication were substantial enough to add noise to the analysis.

Data Categorization

A database was created with an entry for each mean global workload score (n = 1173) reported in each of the 237 included publications. Means reported graphically were estimated. The number of entries per publication ranged from one to 27, with a mean of 4.95, a median of four, and a mode of two entries per publication.

Each entry was categorized by task type. However, this categorization was not straight forward. First, much of the research consisted of multi-task situations. In these cases, the entry was categorized based on the task deemed primary in the publication. Second, for the analysis to have validity, a large number of data points were necessary for each task. In some cases, this was quite easy (e.g., robot operation had 167 entries). However, for other tasks (e.g., crane operation, proof-reading) there were only a couple of entries. The latter were grouped with similar task labels to create “task types.” The 20 task types are presented in Table 3.

Each entry was also categorized by the workload calculation method (i.e., WWL, RTLX, or unknown). Of the 237 publications, 59 reported the WWL for a total of 312 entries. A total of 33 publications reported the RTLX for a total of 154 entries. The majority of publications (n = 145), representing 707 entries, did not report the method of calculating the global workload score. The unknown method was included as it was assumed that one of the two standard weighting methods was utilized unless otherwise reported. This assumption was based on the number of modifications that were reported generally as well as specifically with the weighting method. The ranges and means for the different calculation methods can be seen in Table 4.

A one-way Analysis of Variance (ANOVA) was calculated to determine if significant differences between the unknown weighting method, the RTLX method, and WWL method were indicated. The results, F (2, 1170) = 2.698, p = .068, suggest a failure to reject the null hypothesis. Although, this does not prove that there were not significant differences attributed to weighting method, it supports the decision to present percentile ranks regardless of weighting method.

Table 3
Description of Task Types

Task Type	Description
Air Traffic Control	Real or simulated monitoring and maintenance of safe air space
Card Sorting	Grouping playing cards according to suit in time to a metronome
Classification	Grouping stimuli by shared qualities
Cognitive Activities	Tasks requiring mental action, e.g., computer programming, flight planning, proof-reading, speech shadowing, etc…
Command & Control	Military planning, computer based military simulations, gunner exercises
Computer Activities	Using a computer to balance a checkbook, read email, enter data, …
Daily Activities	Engage in conversation, complete telephone inquiry, and use home medical device
Driving Car	Real or simulated control and operation of motor vehicles
Mechanical Tasks	Assembly tasks, crane operation, and mechanical maintenance
Medical	Emergency room doctor, emergency medical technician, and simulated endoscopic surgery
Memory Tasks	Recall/recollection of stimuli
Monitoring Tasks	Change detection, speech detection, & vigilance
Navigation Tasks	Planning and following a route
Physical Activities	Walking a designated route; dismounted military & police exercises
Pilot Aircraft	Real or simulated control and operation of airplanes/helicopters
Process Control	Real or simulated operation of an engineering system, e.g., power plant
Robot Operation	Real or simulated control of unmanned system
Tracking Tasks	Following a moving stimuli
Video Game	Tetris^{TM}, M-SWAP, etc…
Visual Search Tasks	Active scan of environment for particular stimuli

Table 4
Ranges and Means by Method of Calculating NASA-TLX Global Workload Score

Calculation Method	Min	Max	Mean (SD)
Weighted (WWL)	8.00	80.00	48.74 (14.88)
Unweighted (RTLX)	14.08	88.50	45.29 (14.99)
Unknown	6.21	84.30	48.37 (15.82)

Results

The overall range of scores observed in the literature was 6.21 to 88.5. The deciles and quartiles for all 1173 observations are presented in Table 5. The range and quartiles of observations for each task type are presented in Table 6. The two data points for "No Task" are 12.0 and 14.8. Of note, half of the observed global workloads are between 36.77 and 60.00. Yet, the majority of workload scores observed for daily activities, card sorting, and mechanical tasks were below 36.77.

Table 5
The Deciles and Quartiles of Global NASA-TLX Analysis

Percentile	Score
Min	6.21
10%	26.08
20%	33.00
25%	36.77
30%	39.45
40%	45.00
50%	49.93
60%	53.97
70%	58.00
75%	60.00
80%	62.00
90%	68.00
Max	88.50

A one-way ANOVA was calculated supporting the hypothesis that there are statistically significant differences among the 20 task types, F (19, 1151) = 16.466, p <.001. A post hoc Tukey HSD test was calculated to determine which task types were significantly different from each other. Daily activities, card sorting, and mechanical tasks were significantly different than the majority of other task types. There were many other statistically significant differences among the task types as well. In fact, seven clusters were identified by the Tukey HSD, but none of these clusters were statistically different from the other clusters.

Discussion

Despite the fact that the NASA-TLX has been the most commonly utilized workload measure over the past 25 years, global workload analyses have been limited to comparisons within the same test. This is because no guidance on the interpretation of NASA-TLX scores has been published (Hart, 2006). Specifically, there is no reference a practitioner could cite stating whether an observed workload score was high or low. The purpose of this study was to improve the interpretability of the NASA-TLX global workload scores by providing a description of published scores.

Table 6
Cumulative Frequency Distributions of NASA-TLX Global Workload Scores by Task Type

Task (n)	Min	25%	50%	75%	Max
Air Traffic Control (24)	6.21	42.81	52.44	68.32	85.00
Card Sorting (12)	16.00	21.24	25.63	27.88	49.80
Classification (45)	8.00	30.15	46.00	51.20	84.30
Command & Control (107)	20.00	38.40	50.55	59.50	75.80
Cognitive Tasks (31)	13.08	38.00	46.00	54.66	64.90
Computer Activities (37)	7.46	20.99	54.00	60.00	78.00
Daily Activities (33)	7.20	12.70	18.30	25.90	37.70
Driving Car (36)	15.00	28.05	41.52	51.73	68.50
Medical (45)	9.00	39.35	50.60	61.45	77.35
Mechanical Tasks (22)	20.10	24.90	27.95	33.68	51.03
Memory (37)	6.59	32.62	44.59	66.58	83.50
Monitoring (174)	20.00	39.97	52.24	62.63	77.00
Navigation (14)	19.72	26.35	37.70	52.74	68.90
Physical Activities (21)	40.83	50.98	62.00	71.83	75.19
Pilot Aircraft (152)	16.00	37.70	47.78	54.80	74.00
Process Control (38)	23.90	31.91	42.00	51.83	69.70
Robot Operation (167)	9.59	41.00	56.00	63.00	80.00
Tracking (70)	19.08	39.25	51.00	62.43	88.50
Video Game (60)	14.08	48.23	56.50	63.72	78.00
Visual Search (46)	28.98	51.06	57.89	67.74	79.23

One of the most significant findings of this study is that the observed range is between 6.21 and 88.50. Further, the majority of scores reported (i.e., 80%) were between 26.08 and 68.00. Thus, the range of likely scores appears to be much more restricted than the potential scores. Though, there is great utility from better understanding the range of a wide variety of tasks, it was assumed that practitioners would want the ranges and cumulative frequencies for the 20 different task types.

An auxiliary ANOVA found statistically significant differences between task types. This supports the presentation of cumulative frequencies by task type. However, the large range of scores and the complex pattern of results supported by the post hoc Tukey HSD suggest that the task type is not the only significant factor affecting workload. Other contextual variables (e.g., within task type differences in difficulty, individual ability levels, stress) affect workload as well.

Furthermore, although one can conclude task is an important factor to be considered, these analyses are not sufficient to support authoritative statements as to workload differences among specific task types. Most notably, the task types were broadly defined. Many of the publications examined workload in multi-tasking situations (e.g., Chen and Terrence (2008) examined the combination of a robot operation with a command and control task). These multi-tasking studies were categorized based on primary tasks exclusively. In addition, to have enough data for analysis some related but different tasks (e.g., email, balancing checkbook, etc) were grouped together into task types (e.g., computer activities).

This study does support practitioners by improving the interpretability of NASA-TLX scores. Let us consider a case of a hypothetical new information system for cars. The mean workload of drivers utilizing this system is found to be 58 (+/-4). Considering the potential range of scores on the NASA-TLX, this is above the midpoint but not remarkably so. However considering the frequency distribution of scores in this study for all tasks, the mean 58 is higher than 70% of the scores. Even the lower end of the confidence interval, 54, is higher than 60% of all scores. Moreover, if the practitioner considers only the scores obtained for driving tasks 54 is well above 75% of the observed scores. The practitioner can safely say that high workload was observed in the test.

Let us further assume that the practitioner's job is to state if the workload is acceptable or unacceptable; the latter of which would require system redesign. Such a decision requires considering the context of the test. If the test is conducted with novice drivers at night on unfamiliar roads this leads the practitioner to a different conclusion than if the participants were experienced drivers during daylight on familiar streets. In the former situation, assuming performance was adequate for both driving and interaction with the information/comfort system, the high workload is anticipated and therefore may be deemed acceptable. Conversely, in the latter situation, the high workload will most likely be unanticipated and therefore deemed unacceptable and a system redesign is therefore required.

Without these scores, the practitioner could only state that the mean workload was 58 (+/- 4). To indicate that this workload was high, she would have needed to do a comparison test with other systems or other conditions. A comparison test would have required more resources to conduct.

It should be noted that this research was not intended to contribute to the discussion of "red-lines." A "red-line" would be a point at which performance degrades significantly (Grier et al., 2008). This study did not consider participant performance. The relationship between workload and performance is complicated (Gopher & Donchin, 1986; Muckler & Seven, 1992). Performance can degrade when workload is excessive. However, there is a period where workload is high, but does not result in performance degradation. As, Muckler and Seven (1992, p. 449) wrote "the operator's awareness of increasing effort being used, even before any performance degradation occurs, should give (survey based) measures a special role to play. A person can integrate diverse considerations and predict their future impact."

Performance also degrades when there are not sufficient demands, but this is not associated with workload. A low demand situation leads to complacency or the vigilance paradigm, both of which result in performance degradation (Grier, 2011; Grier et al., 2003). Complacency is a "self-satisfaction, which may result in non-vigilance based on an unjustified assumption of satisfactory system state" (Billings, Lauber, Funkhouser, Lyman, and Huff, 1976). The vigilance paradigm is a low demand task in which people are asked to sustain attention and respond to a rare critical signal. Of these two low demand situations, only complacency is associated with low workload (Grier, 2011). Vigilance is actually associated with high workload (Grier et al, 2003).

The findings of complacency and vigilance research support the idea that the experience of workload is more than just the mental, physical, and temporal demands of the task. The individual's perception of their performance, their frustration, and their effort to perform the task also contribute to workload as assessed by the NASA-TLX. This idea is further supported by this analysis.

Two NASA-TLX scores of 12.0 and 14.8 were reported in studies in which participants did nothing but wait. If demands were the only relevant component of workload, one would expect these two scores to be the lowest observed workload scores. This was not the case. The lowest workload reported was 6.21 for an air traffic control task. Furthermore, 21 entries were less than 12 (1.8% of all entries) and a further 11 entries (.9% of entries) were less than 14.8.

In conclusion, this study represents a first step to improving the interpretability of NASA-TLX scores. This study provides ranges and percentile ranks for practitioners to use to determine how common a particular workload score is. In other words the practitioner can state if this score is high or low in comparison to other studies of similar tasks. To determine if the workload is acceptable, the practitioner must consider not only the workload score, but also the different contextual variables (e.g., task type, different levels of expertise, different levels of difficulty within task type, different stressors), and the human-system performance.

References

Billings, C. E., Lauber, J. K., Funkhouser, H., Lyman, G. and Huff, E. M. (1976). *NASA aviation safety reporting system* (No. TM-X-3445). Moffet Field, CA: NASA Ames Research Center.

Bittner, A. C., 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 Human Factors and Ergonomics Society Annual Meeting*, 33, 1476-80.

Byers, J. C., Bittner, A. C., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 32, 1145-49.

Chen, J. Y., & Terrence, P. I. (2008). Individual differences in concurrent performance of military and robotics tasks with tactile cueing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52, 1407-1411.

Gopher, D., and Donchin, E. (1986) Workload: An examination of the concept. In Boff, K.R., Kaufmann, L., & Thomas, J.P. (Eds). *Handbook of Perception and Human Performance, Volume 2,* Wiley & Sons, New York.

Grant, R., Carswell, C. M., Lio, C., Seales, W. B., & Clarke, D. (2008). Equivalent-forms reliability of printed and spoken versions of the NASA-TLX. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52, 1532-35.

Grier, R. A. (2011). Cognitive readiness at the tactical level: A review of measures. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 55, 404-8. doi: 10.1177/1071181311551083

Grier, R. A., Warm, J. S., Dember, W. N., Matthews, G., Galinsky, T. L., Szalma, J. L., & Parasuraman, R. (2003). The vigilance decrement reflects limitations in effortful attention, not mindlessness. *Human Factors*, 45, 349-359.

Grier, R., Wickens, C., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., & John, M. S. (2008). The red-line of workload: Theory, research, and design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 52, 1204-08.

Hart, S. G. (2006). NASA-task load index (NASA-TLX): 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50, 904-8.

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* (139-183). Amsterdam: North Holland Press.

Moroney, B. W., Warm, J. S., & Dember, W. N. (1995). Effects of demand transitions on vigilance performance and perceived workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 39, 1375-79.

Muckler, F. A., & Seven, S. A. (1992). Selecting performance measures:" Objective" versus" subjective" measurement. *Human Factors*, 34(4), 441-455.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors*, 33(1), 17-33.

Sharek, D. (2011). A useable, online NASA-TLX tool. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 55, 1375-79.