Gawron, V.J. (2000)

Human Performance Measures Handbook.
Mahwaw, NJ: Lawrence Erlbaum Associates

3.2.23 NASA Task Load Index

General description - The NASA Task Load Index (TLX) is a multi-dimensional subjective workload rating technique (see Figure 15). In TLX, workload is defined as the "cost incurred by human operators to achieve a specific level of performance." The subjective experience of workload is defined as an integration of weighted subjective responses (emotional, cognitive, and physical) and weighted evaluation of behaviors. The behaviors and subjective responses, in turn, are driven by perceptions of task demand. Task demands can be objectively quantified in terms of magnitude and importance. An experimentally based process of elimination led to the identification of six dimensions for the subjective experience of workload: mental demand, physical demand, temporal demand, perceived performance, effort, and frustration level. The rating-scale definitions are presented in Table 24.

Strengths and limitations - Sixteen investigations were carried out, establishing a database of 3461 entries from 247 subjects (Hart and Staveland, 1987). All dimensions were rated on bipolar scales ranging from 1 to 100, anchored at each end with a single adjective. An overall workload rating was determined from a weighted combination of scores on the six dimensions. The weights were determined from a set of relevance ratings provided by the subjects.

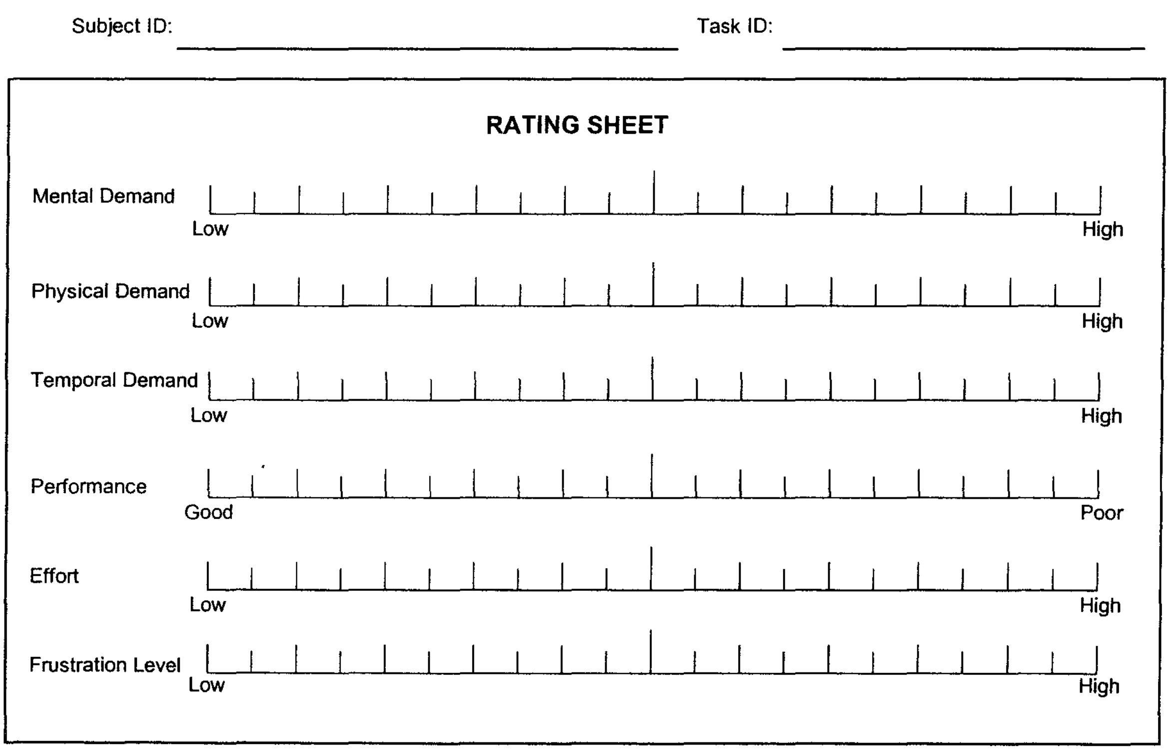

FIG. 15. NASA TLX Rating Sheet

TABLE 24.
NASA TLX Rating-Scale Descriptions

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? (NASA Task Load Index, p. 13)

Hart and Staveland (1987) concluded that the TLX provides a sensitive indicator of overall workload as it differed among tasks of various cognitive and physical demands. They also stated that the weights and magnitudes determined for each TLX dimension provide important diagnostic information about the sources of loading within a task. They reported that the six TLX ratings took less than a minute to acquire and suggested the scale would be useful in operational environments.

Battiste and Bortolussi (1988) reported significant workload effects as well as a test-retest correlation of +0.769. Corwin, Sandry-Garza, Biferno, Boucek, Logan, Jonsson, and Metalis (1989) reported that NASA TLX was a valid and reliable measure of workload.

NASA TLX has been used extensively in the flight environment. Bittner, Byers, Hill, Zaklad, and Christ (1989), Byers, Bittner, Hill, Zaklad, and Christ (1988), Hill, Byers, Zaklad, and Christ (1989), Hill, Zaklad, Bittner, Byers, and Christ (1988), and Shively, Battiste, Matsumoto, Pepitone, Bortolussi, and Hart (1987), based on in-flight data, stated that TLX ratings significantly discriminated flight segments.

Nataupsky and Abbott (1987) successfully applied NASA TLX to a multi-task environment. Vidulich and Bortolussi (1988) replicated the significant flight-segment effect but reported no significant differences in TLX ratings between control configurations or between combat countermeasure conditions. In a later study, Tsang and Johnson (1989) reported reliable increases in NASA TLX ratings when target-acquisition and engine-failure tasks were added to the primary flight task. Vidulich and Bortolussi (1988) reported significant increases in NASA TLX ratings from the cruise to the combat phase during a simulated helicopter mission. There was no affect of control configuration, however.

Nygren (1991) reported that NASA TLX is a measure of general workload experienced by aircrews. Selcon, Taylor, and Koritsas (1991) concluded from pilot ratings of an air combat flight simulation that NASA TLX was sensitive to difficulty but not the pilot experience. Hancock, Williams, Manning, and Miyake (1995) reported the NASA TLX score was highly correlated with difficulty of a simulated flight task.

Hendy, Hamilton, and Landry (1993) examined undimensional and multidimensional measures of workload in a series of four experiments (low-level helicopter operations, peripheral version display evaluation, flight simulator fidelity, and aircraft landing task). They concluded that if an overall measure of workload is required, then a univariate measure is as sensitive as an estimate derived from multivariate data. If a univariate measure is not available then a simple unweighted additive method can be used to combine ratings into an overall workload estimate.

Byers, Bittner, and Hill (1989) suggested using raw TLX scores. Moroney, Biers, Eggemeier, and Mitchell (1992) reported that the pre-rating weighting scheme is unnecessary since the correlation between weighted and unweighted scores was +0.94. Further, delays of 15 minutes did not affect the workload ratings; delays of 48 hours, however, did. After 48 hours, ratings no longer discriminate workload conditions. Moroney, Biers, and Eggemeier (1995) concluded from a review of relevant studies that 15-minute delays do not affect NASA TLX.

TLX has been applied to other environments. Hill, Iavecchia, Byers, Bittner, Zaklad, and Christ (1992) reported that the NASA TLX was sensitive to different levels of workload and high in user acceptance. Their subjects were Army operators. Jordan and Johnson (1993) concluded from an on-road evaluation of a car stereo that TLX was a useful measure of mental workload.

Hancock and Caird (1993) reported a significant increase in the overall workload rating scale of the NASA TLX as shrink rate of a target decreased. The highest ratings were on paths with four steps rather than 2, 8, or 16 steps from cursor to target. NASA TLX scores significantly increased as ambient noise increased (Becker, Warm, Dember, and Hancock, 1995).

Harris, Hancock, Arthur, and Caird (1995) reported significantly higher ratings on five (mental demand, temporal demand, effort, frustration, and physical demand) of the six NASA-TLX scales for manual than automatic tracking.

Vidulich and Tsang (1985) compared the SWAT and TLX. They stated that the collection of ratings is simpler with SWAT. However, the SWAT card sort is more tedious and time consuming. Battiste and Bortolussi (1988) reported no significant correlation between SWAT and NASA TLX in a simulated B-727 flight. Hancock (1996) stated that NASA TLX and SWAT "were essentially equivalent in terms of their sensitivity to task manipulations." The task was tracking. Tsang and Johnson (1987) reported good correlations between NASA TLX and a one dimensional workload scale. Vidulich and Tsang (1987) replicated the Tsang and Johnson finding as well as reported a good correlation between NASA TLX and the Analytical Hierarchy Process.

Vidulich and Pandit (1987) reported only three significant correlations between NASA TLX and seven personality tests (Jenkins Activity Survey, Rotter's Locus of Control, Cognitive Failures Questionnaire, Cognitive Interference Questionnaire, Thought Occurrence Questionnaire, California Q-Sort, and the Myers-Briggs Type Indicator): the speed scale of the Jenkins Activity Survey and the physical demand scale of the NASA TLX ($r = -0.23$), Locus of control and physical demand ($r = +0.21$), and finally locus of control and effort ($r = +0.23$).

Data requirements - Use of the TLX requires two steps. First, subjects rate each task performed on each of the six subscales. Hart suggests that subjects should practice using the rating scales in a training session. Second, subjects must perform 15 pair-wise comparisons of six workload scales. The number of times each scale is rated as contributing more to the workload of a task is used as the weight for that scale. Separate weights should be derived for diverse tasks; the same weights can be used for similar tasks. Note that a set of IBM PC compatible programs has been written to gather ratings and weights and to compute the weighted workload scores. The programs are available from the Human Factors Division at NASA Ames Research Center, Moffett Field, CA.

Thresholds - Knapp and Hall (1990) used NASA TLX to evaluate a highly automated communication system. Using 40 as a high workload threshold, the system was judged to impose high workload and difficult cognitive effort on operators.

Sources -

Battiste, V. and Bortolussi, M.R. Transport pilot workload: a comparison of two objective techniques. Proceedings of the Human Factors Society 32nd Annual Meeting. 150-154; 1988.

Becker, A.B., Warm, J.S., Dember, W.N., and Hancock, P.A. Effects of jet engine noise and performance feedback on perceived workload in a monitoring task. International Journal of Aviation Psychology. 5(1), 49-62, 1995.

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

Byers, J.C., Bittner, A.C., and Hill, S.G. Traditional and raw Task Load Index (TLX) correlations: are paired comparisons necessary? In Advances in industrial erogonomics and safety. London: Taylor and Frances; 1989.

Byers, J.C., Bittner, A.C., Hill, S.G., Zaklad, A.L., and Christ, R.E. Workload assessment of a remotely piloted vehicle (RPV) system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1145-1149). Santa Monica, CA: Human Factors Society; 1988.

Corwin, W.H., Sandry-Garza D.L., Biferno, M.H., Boucek, G.P., Logan, A.L., Jonsson, J.E., and Metalis, S.A. Assessment of crew workload measurement methods, techniques, and procedures. Volume I-Process, methods, and results (WRDC-TR-89-7006). Wright-Patterson Air Force Base, OH; 1989.

Hancock, P.A. Effects of control order, augmented feedback, input device, and practice on tracking performance and perceived workload. Ergonomics. 39(9), 1146-1162, 1996.

Hancock, P.A. and Caird, J.K. Experimental evaluation of a model of mental workload. Human Factors. 35(3), 413-419; 1993.

Hancock, P.A., William G., Manning, C.M., and Miyake, S. Influence of task demand characteristics on workload and performance. International Journal of Aviation Psychology. 5(1), 63-86, 1995.

Harris, W.C., Hancock, P.A., Arthur, E.J., and Caird, J.K. Performance, workload, and fatigue changes associated with automation. International Journal of Aviation Psychology. 5(2), 169-185; 1995.

Hart, S.G. and Staveland, L.E. 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: Elsevier; 1987.

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

Hill, S.G., Byers, J.C., Zaklad, A.L., and Christ, R.E. Subjective workload assessment during 48 continuous hours of LOS-F-H operations. Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 1129-1133). Santa Monica, CA: Human Factors Society; 1989.

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklad, A.L., and Christ, R.E. Comparison of four subjective workload rating scales. Human Factors. 34:429-439; 1992.

Hill, S.G., Zaklad, A.L., Bittner, A.C., Byers, J.C., and Christ, R.E. Workload assessment of a mobile air defense system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1068-1072). Santa Monica, CA: Human Factors Society; 1988.

Jordan, P.W. and Johnson, G.L. Exploring mental workload via TLX: the case of operating a car stereo whilst driving. In A.G. Gale, I.D. Brown, C.M. Haslegrave, H.W. Kruysse, and S.P. Taylor (Eds.). Vision in vehicles - IV. Amsterdam: North-Holland; 1993.

Knapp, B.G. and Hall, B.J. High performance concerns for the TRACKWOLF system (ARI Research Note 91-14). Alexandria, VA; 1990.

Moroney, W.F., Biers, D.W., and Eggemeier, F.T. Some measurement and methodological considerations in the application of subjective workload measurement techniques. International Journal of Aviation Psychology. 5(1), 87-106, 1995.

Moroney, W.E., Biers, D.W., Eggemeier, F.T., and Mitchell, J.A. A comparison of two scoring procedures with the NASA Task Load Index in a simulated flight tasks. NAECON Proceedings (pp. 734-740). Dayton, OH; 1992.

Nataupsky, M. and Abbott, T.S. Comparison of workload measures on computer-generated primary flight displays. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 548-552). Santa Monica, CA: Human Factors Society; 1987.

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

Selcon, S.J., Taylor, R.M., and Koritsas, E. Workload or situational awareness?: TLX vs. SART for aerospace systems design evaluation. Proceedings of the Human Factors Society 35th Annual Meeting. 62-66, 1991.

Shively, R.J., Battiste, V., Matsumoto, J.H., Pepitone, D.D., Bortolussi, M.R., and Hart, S.G. Inflight evaluation of pilot workload measures for rotorcraft research. In R.S. Jensen Proceedings of the 4th Symposium on Aviation Psychology (pp. 637-643). Columbus, OH: Ohio State University; 1987.

Tsang, P.S. and Johnson, W. Automation: Changes in cognitive demands and mental workload. Proceedings of the Fourth Symposium on Aviation Psychology. Columbus, OH: Ohio State University; 1987.

Tsang, P.S. and Johnson, W.W. Cognitive demands in automation. Aviation, Space, and Environmental Medicine. 60, 130-135; 1989.

Vidulich, M.A. and Bortolussi, M.R. Control configuration study. Proceedings of the American Helicopter Society National Specialist's Meeting: Automation Application for Rotorcraft; 1988.

Vidulich, M.A. and Bortolussi, M.R. Speech recognition in advanced rotorcraft: Using speech controls to reduce manual control overload. Proceedings of the National Specialists' Meeting Automation Applications for Rotorcraft, 1988.

Vidulich, M.A. and Pandit, P. Individual differences and subjective workload assessment: Comparing pilots to nonpilots. Proceedings of the International Symposium on Aviation Psychology. 630-636, 1987.

Vidulich, M.A. and Tsang, P.S. 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; 1985.

Vidulich, M.A. and Tsang, P.S. Absolute magnitude estimation and relative judgment approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society; 1987.