

THE TLX: ONE OR MORE CONSTRUCTS

L. L. Bailey, Ph.D., R. C. Thompson, Ph.D.
Federal Aviation Administration
Civil Aeromedical Institute, Oklahoma City, OK

ABSTRACT

The NASA Task Load Index (TLX) is a six-item scale reported to measure six facets of subjective workload. These include mental demand, physical demand, temporal demand, self-assessment of performance, amount of effort, and the amount of frustration. In this study, we examined the internal consistency and factor structure of the TLX to determine whether respondents perceive the TLX items to be separate facets or one unifying construct. To determine the internal consistency of the TLX, data from five studies representing three environments (two laboratory, one classroom, and two organizations) were analyzed separately. The data were then collapsed across the five studies to determine an overall factor structure. Measures of internal consistency suggested that respondents' perceive the TLX not to be measuring six separate facets, but instead to be measuring one and perhaps two constructs. These results were further supported from a subsequent factor analysis. The underlying theme of the first construct (consisting of five items) was related to time pressure. A single item, the self-assessment of performance, represented a second construct. Based on these findings it is recommended that caution be used when interpreting individual TLX items.

INTRODUCTION

Since its inception, human factors practitioners have used the National Aeronautic and Space Administration Task Load Index (TLX; Hart & Staveland, 1988) to measure operators' subjective workload. Although numerous studies have demonstrated substantial TLX correlations with objective performance indicators (i.e., predictive validity), relatively few studies have examined the psychometric properties of the TLX (Eggemeier, Wilson, Kramer, & Damos, 1991; Nygren, 1991). This is true not only for the TLX but for subject workload measures as a whole. In particular, subjective workload measures are commonly used without a clear understanding of what component or attribute of workload is being measured (Nygren, 1991). For example, does the TLX measure the

cognitive component of workload, the affective component, or both? One reason that it is difficult to answer this question is the manner in which the TLX was developed.

The TLX was constructed using what is referred to as the external criterion reference strategy (Burisch, 1984; Comrey, 1988). In this strategy, a pool of items are developed based on a content sampling of the domain of interest. Item responses are then obtained and correlated with the criteria of interest (in this case, measures of objective workload) to identify those items with the highest statistically significant correlations. The selected items are then examined for colinearity to eliminate item redundancy. Although scales developed in this way may have predictive validity, the individual items comprising the final scale may not adequately represent the construct they were intended to measure. As Comrey (1988) notes: "...scales developed by the criterion-group method, despite their possible practical utility, are not good marker variables for factor constructs because they are factorially complex by virtue of their method of construction" (p. 756).

In contrast to the external criterion approach, the factor analytic approach to scale development begins by defining the various second order constructs (e.g., mental demand and physical effort) that are subsumed under the construct of interest (in this case subjective workload). A set of items are then developed to represent each second order construct. Once item responses are collected, factor analytic techniques are employed to identify a subset of items to serve as marker variables for their respective second order constructs. Although scales developed in this way may have what is called factor validity, further work is needed to determine the scale's predictive validity.

Both methods of scale development have their strengths and weaknesses. Rather than choosing one method over the other, a mixed approach to scale development might better serve the practitioner and scientific community. This would involve viewing scale development as a process that involves establishing both its practical utility and construct validity. Since the TLX has already established its

practical utility by its relationship with objective measures of operator workload, what remains to be accomplished is to examine the nature of the construct the TLX purports to measure.

In its current form, the TLX purports to measure six facets of subject workload, each of which is represented by a single item. By measuring six facets of subjective workload the developers hoped to provide diagnostic features within their scale. However, it is one thing to create items with the face validity that they are measuring different facets, and it is something altogether different to empirically demonstrate that separate facets are in fact being measured. In this report, we examine the factor structure and the internal consistency of the TLX to determine if the pattern of TLX ratings indicates the presence of a single or multifaceted construct.

METHOD

Review of Archival Data

Using archival data from five studies, the psychometric properties of the TLX were examined across three environments. Two data sets were obtained from *laboratory* experiments, one data set was obtained from a *classroom* setting, and two data sets were obtained from *organizational* settings. The five studies are summarized below.

The two laboratory studies were conducted using a high psychological fidelity, multi-sector, personal computer-based, ATC team training device that simulated radar-based air traffic control (ATC) tasks (Bailey, Broach, Thompson, & Enos, 1999; Bailey & Thompson, 2000). In this environment, airspace configuration and traffic were controlled to assess how system effectiveness measures (safety errors, aircraft delay time, and proportion of aircraft reaching their destination) changed as a function of training and aircraft density. One hundred twenty high school graduates (ages 18-30, 63% male) participated in the first study, and 240 high school graduates (ages 18-30, 47% male) participated in the second study. Participants were recruited through a local temporary help provider and received payment for their participation. In both studies, the TLX was administered immediately following performance in a 28-minute ATC scenario.

Thirty-two entry-level ATC specialists participated in classroom exercises on teamwork that used the same simulated radar-based ATC training device employed in the previous laboratory studies. No age

or gender data were available. Again, TLX data was administered immediately following performance of a 28-minute training exercise.

In the first organizational study, 70 out of 133 employees (53% response rate) from the Civil Aviation Registry participated in an Organizational Assessment Survey (OAS) (Worley, Bailey, Thompson, Joseph, & Williams, 1999). Embedded within the OAS was the TLX. Participants (83% female, median age = 49) completed the survey in a group setting during a one hour block of time. To ensure the participant's anonymity, no employee identification was included on the survey. Before completing the TLX, participants were asked to reflect on their work over the past 30 days. It should be noted that a 30-day time frame represents a non normal use of the TLX. Normally the TLX is administered either during or immediately following performance.

The second organizational study involved 54 employees of the United States Coast Guard who participated in an occupational fatigue study conducted by the FAA. No age or gender data were available. As part of the study's protocol, the participants' reaction times were recorded after varying times of an eight-hour day while performing a psychomotor vigilance test. Prior to testing, participants were asked to reflect on their current work and complete the TLX.

Measures

Subjective workload was assessed in all studies using the TLX. In the TLX, subjective workload is viewed as a multidimensional construct involving the subjective appraisal of one's: (1) mental demand, (2) physical demand, (3) temporal demand, (4) performance, (5) effort, and (6) frustration level. These dimensions were defined and presented as single items in a questionnaire format as suggested by Nygren (1991). Participants used a 21-point scale (1 = low/good, 21 = high/poor) to indicate their responses.

Procedures

TLX data were collected from each study and scored so that lower scores reflected less workload or more favorable performance. Each study was analyzed separately on measures of internal consistency and then aggregated to determine an overall factor structure. All analyses were performed using the Statistical Package for the Social Sciences (SPSS), version 9.0.

RESULTS

Two measures of internal consistency were examined: (1) item-total correlations and (2) Cronbach's alpha (α). The item-total correlation process involves the removal of one item from the scale. A correlation coefficient is then computed between the removed item and the linear composite of the remaining items. This procedure is performed for each item. The net result is a correlation coefficient derived for each item in the scale. Items with the highest item-total correlations are thought to be the primary factors governing the remaining inter-item correlations. As a rule of thumb, item-total correlation coefficients less than .30 indicate that an item is not accounting for much variance in the overall scale. Item-total correlations in excess of .70 generally indicate a given item is a redundant measure of the rest of the scale items.

In addition to item-total correlations, Cronbach's alpha provides a summary statistic of the overall response pattern of scale items. The statistic is based on the premise that items comprising a single construct should evoke similar response patterns. That is, respondents will consistently rate items representing the same construct in the same way (e.g., scoring all ones, or all threes, or all fives on a 1-5 rating scale). Cronbach's alpha ranges from $\alpha = 0$ (indicating the absence of a consistent response pattern) to $\alpha = 1$ (indicating a perfect response pattern match). For scales similar to those used in this report, acceptable alphas range between .70 to .90 (Nunnally, 1978). Table 1 shows the TLX item-total correlations and corresponding Cronbach's alpha for laboratory, classroom, and organizational settings.

Of particular interest to this study was the similarity of the rankings of item-total correlations across measurements. Notice that for both laboratory and organizational settings, temporal demand had the highest item total correlation ($r = .71$ to $.91$) and performance had the smallest ($r = .17$ to $.43$). This suggests that in three diverse settings, TLX scores were driven primarily by factors associated with time pressures and least affected by a self-assessment of one's performance.

The last psychometric analysis involved a factor analysis of the TLX items. In factor analysis the correlation matrix is summarized (i.e., reduced) based on the amount of shared variance among the variables. Variables that share the most variance are grouped into a domain called a factor. Multiple factors are an indication of multiple constructs that

are independent from one another. One way of evaluating the results of a factor analysis is to examine the factor loadings associated with each variable. Factor loadings range from -1.0 to +1.0 and are similar to a regression beta weight in that they determine the strength of the association between a given variable and the construct represented by the factor. Factor loadings below $|\ .30|$ are typically considered to be a weak association.

Table 1

TLX, Item-Total Correlations and Coefficient Alpha

TLX Item	Item-Total Correlations				
	Lab S1*	Lab S2	Class S3	Org S4	Org S5
Mental	.77	.68	.84	.78	.71
Physical	.51	.38	.85	.45	.26
Temporal	.84	.69	.91	.83	.56
Performance	.17	.27	.43	.22	.20
Effort	.68	.68	.75	.78	.61
Frustration	.77	.58	.67	.67	.41
Sample Size	120	240	32	70	54
Alpha	.79	.78	.90	.81	.74
Alpha without Performance	.86	.80	.93	.86	.76

* S1-S5 = Study 1 - Study 5

Table 2 shows the results of the principal axis factor analysis with a varimax rotation. Notice that two distinct factors emerged accounting for 59% of the total variance. With the exception of performance, all items are strongly associated with the first factor (accounting for 54% of the total variance), with time demands having the highest factor loading (.90). Independent from the first is the second factor (accounting for 5% of the total variance), containing only one item, the self-assessment of one's performance (loading of .68). This suggests that the self-assessment of one's performance is unrelated to the remaining five items of the TLX.

Table 2

TLX Rotated Factor Matrix

TLX Item	Factor Loadings	
	Factor 1	Factor 2
Mental	.85	.00
Physical	.61	-.14
Temporal	.90	.01
Performance	.05	.68
Effort	.85	-.03
Frustration	.77	.10

CONCLUSION

Although the TLX was designed to represent six different facets of subjective workload, the results of this study suggest that, when completing the TLX, respondents perceive only one or, perhaps, two constructs. Because of this, users of the TLX should use caution when analyzing individual items as representing different facets of subjective workload. Five of the six TLX items demonstrated internal consistency, thus suggesting the presence of a single underlying construct associated with temporal demands/time pressure. This may be driven by the fact that time pressure perceptions strongly influence the other components measured by the TLX such that the remaining items are no longer independent measures of a given facet. A sixth item, the self-assessment of one's performance appears to be unrelated to the other five. This raises the question of whether performance self assessment should be retained in the scale. To the extent that these findings are replicated in other studies involving both cognitive and physically demanding tasks, then adjustments to the TLX are warranted. These include: (1) adding additional items so that 6 subscales are created which reflect the original 6 workload facets, (2) creating a self-assessment of one's performance subscale by adding more items related to performance, or (3) dropping the self-assessment of one's performance from the TLX scale.

REFERENCES

Bailey, L., & Thompson, R. (2000). *The Effects of performance feedback on air traffic control team*

coordination: A simulation study. (DOT/FAA/AM-00/25). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Bailey, L., Broach, D., Thompson, R., & Enos, R. (1999) *Controller teamwork evaluation and assessment methodology (CTEAM): A scenario calibration study.* (DOT/FAA/AM-99/24). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Burisch, M. (1984). Approaches to personality inventory construction. *American Psychologist*, 39, 214-27.

Comrey, A. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, 56, 754-61.

Eggemeier, F., Wilson, G., Kramer, A., & Damos, D. (1991). Workload assessment in multi-task environments. In D. Damos (Ed.), *Multiple-task performance* (pp. 207-16). Washington, DC: Taylor & Francis Ltd.

Hart, S., Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. Hancock, & N. Meshkati (Eds.), *Human Mntal Workload* (pp. 139-83). New York: North-Holland

Nunnally, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

Nygren, T. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors*, 33(1), 17-31.

Worley, J., Bailey, L., Thompson, R., Joseph, K., Williams, C. (1999). *Organizational communication and trust in the context of technology change.* (DOT/FAA/AM-99/25). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine