Ergonomics, vol. 64, no. 7 (2021), pp. 869-878
Author version. Published version available at: https://doi.org/10.1080/00140139.2021.1876927

Reference Values and Subscale Patterns for the Task Load Index (TLX): A Meta-Analytic Review

Morten Hertzum

University of Copenhagen, Copenhagen, Denmark, hertzum@hum.ku.dk

Abstract. The Task Load Index (TLX) is the predominant instrument for self-reporting workload. On the basis of a meta-analytic review of 556 studies, this paper supplies reference values for TLX and its six subscales across domains, technologies, regions, and real-life/lab settings. Across domains, TLX spans mean values from 35 for leisure to 56 for manual labor. TLX tends to be driven upward by the subscales of mental demand and effort and downward by the subscales of physical demand and frustration. For technologies, handheld devices are associated with lower TLX, possibly because they are simpler and more task-specific. TLX also varies across regions in that it is higher for studies in Asia than in Europe and North America. This variation is only partly explained by co-variation in domains. Furthermore, TLX is higher and its subscales more inter-correlated when it is studied in real-life rather than lab settings.

Keywords: mental workload, NASA-TLX, task load index, TLX, workload

Practitioner summary: Practitioners can use the reference values supplied in this paper to benchmark their TLX measurements against those from the corpus of TLX research. Furthermore, the reported subscale patterns add to the diagnostic power of the TLX instrument.

1 Introduction

Workload influences the performance and experience of tasks and is, therefore, a central concept in human factors and ergonomics (Epps, 2018; Vidulich & Tsang, 2012; Young et al., 2015). It emerges from the interaction among the demands imposed by a task, the circumstances under which the task is performed, and the skills, behaviors, and perceptions of the person performing the task (Hart & Staveland, 1988). In measuring workload, the task load index (TLX, aka NASA-TLX; Hart & Staveland, 1988) has become a widely accepted instrument.

TLX is a questionnaire instrument for self-reporting workload. Other instruments for measuring workload include the modified Cooper-Harper scale (Wierwille & Casali, 1983), the subjective workload assessment technique (Reid & Nygren, 1988), and the workload profile (Tsang & Velazquez, 1996). However, this study focuses on TLX – because of its widespread use across multiple domains (Grier, 2015; Hart, 2006). Workload measurements complement performance measurements by providing data about how tasks are experienced (de Waard & Lewis-Evans, 2014; Hancock & Matthews, 2019). Such data are important because people who experience the workload of a task as excessive will behave as though they are overloaded, irrespective of the objective task demands (Hart & Staveland, 1988).

The aim of this study is to supply reference values for TLX. Reference values provide information about whether a TLX value is low or high compared to an independent corpus of TLX measurements. Thereby, researchers and practitioners can benchmark the workload imposed by a specific task or technology against others' measurements. To be useful, reference values should take into account the existence of factors that systematically influence workload. Previous studies have suggested that workload varies across domains (Grier, 2015), technologies (Yan et al., 2017), world regions (Johnson & Widyanti, 2011), and real-life/lab settings (Niforatos et al., 2018). To increase the diagnostic power of TLX measurements, previous studies have also recommended that the individual subscales of the TLX should be examined, not just the overall TLX score (Galy et al., 2018). For these reasons, the present study has the additional aim of showing patterns in how TLX and its subscales vary across domains, technologies, regions, and real-life/lab settings. These patterns provide for tailoring the reference values to specific situations. Specifically, the subscale patterns provide for benchmarking individual workload dimensions and for assessing their relative contribution to overall workload.

The reference values supplied in this study are obtained through a meta-analytic review of the TLX values published in the 30-year period 1990-2019. Previous reviews of TLX have not reported TLX values at all (Hart, 2006) or only reported values for the overall TLX scale (Grier, 2015). The need for reference values for the subscales is illustrated by the studies that focus on the subscales to the extent of not even reporting the overall TLX (e.g., Bliss & Hanson, 2018; Krekhov & Krüger, 2019). These studies use the TLX instrument as an inventory of workload dimensions that can be analyzed individually and – when all subscales are analyzed – span the complete experience of workload.

2 The TLX instrument

Hart and Staveland (1988) developed and validated TLX in a multi-year project at the National Aeronautics and Space Administration (NASA). Developed at NASA, the instrument is also known as NASA-TLX. It has become so widely used that de Winter (2014, p. 293) stated that "workload has become synonymous with the TLX". The instrument consists of six subscales that measure somewhat independent dimensions of workload (Hart & Staveland, 1988):

- *Mental demand* (MD): "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* (PD): "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* (TD): "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* (EF): "How hard did you have to work (mentally and physically) to accomplish your level of performance?"
- *Performance* (PE): "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* (FR): "How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?"

Each subscale is measured with a single item. Hence, the entire instrument consists of six items. The item endpoints are 'Low' and 'High', except for PE, which has the endpoints 'Good' and 'Poor'. The small number of items makes the instrument easy to administer. At the same time, the individual subscales can be reported along with the aggregate TLX value, thereby increasing the diagnostic power of the instrument. Often, TLX is simply calculated as the mean of the six item ratings. When calculated in this way, TLX is sometimes referred to as raw TLX (Hart, 2006).

Raw TLX stands in contrast to weighted TLX. Originally, Hart and Staveland (1988) specified that the rating of the six subscales should be followed by a weighting procedure, which consists of indicating the more significant subscale in each of the 15 possible pairs of subscales. The weight of each subscale is the number of times it is deemed the more significant. TLX is then calculated as the sum of the weighted contribution (i.e., rating times weight) of each subscale divided by 15. The weighting is intended to tailor the TLX instrument to the task by emphasizing the most significant dimensions of the task. However, the weighting procedure has been depreciated because it has been found to be ineffective (Hendy et al., 1993; Nygren, 1991).

3 Method

Following procedures for systematic reviews, 556 studies were selected and analyzed.

3.1 Inclusion criteria

The authoritative reference for TLX is inarguably Hart and Staveland (1988). Thus, the primary inclusion criterion for this review was that papers had to cite Hart and Staveland (1988). This criterion ensured that all included papers defined TLX in the same way. In addition, the papers had to report empirically obtained values for all six subscales. Conversely, papers were excluded if they were not based on empirical data or only reported an aggregate TLX value. A few additional criteria served to bolster the quality of the data set. In total, each paper had to satisfy seven inclusion criteria:

- Papers that cited Hart and Staveland (1988)
- Papers published in the 30-year period 1990-2019
- Empirical studies with at least five participants
- Papers that reported values for all six TLX subscales
- Research papers published in journals, edited books, and conference proceedings
- Only the most extensive paper when multiple versions existed
- Papers in English

3.2 Paper-selection process

The paper-selection process involved multiple steps, see Figure 1. First, Google Scholar was searched for the papers that cited Hart and Staveland (1988), were published in the period 1990-2019, and contained the terms mental demand, physical demand, temporal demand, effort, performance, and frustration (i.e., the six subscales).

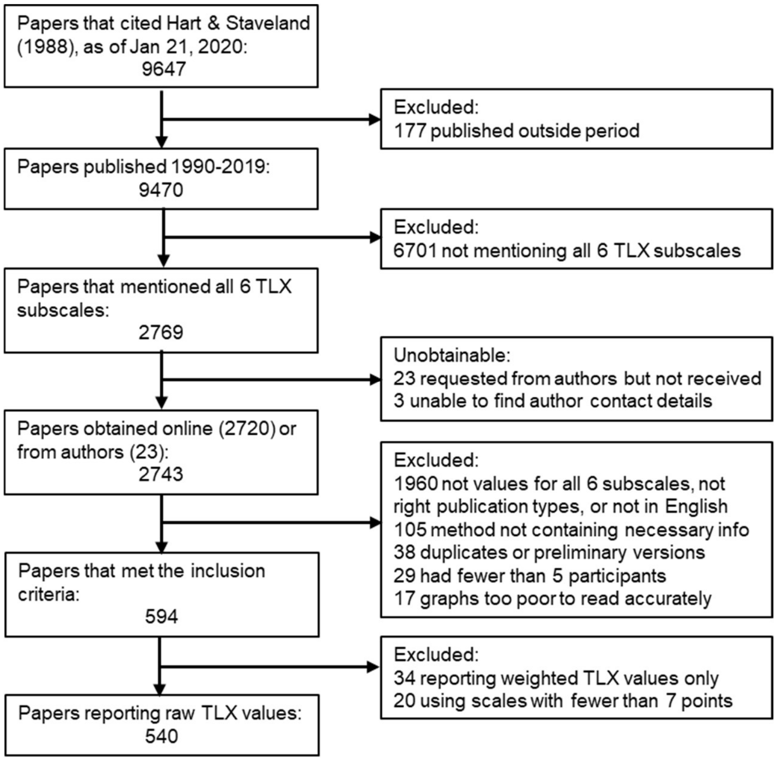
Papers that cited Hart & Staveland (1988), as of Jan 21, 2020: 9647
Excluded: 177 published outside period
Papers published 1990-2019: 9470
Excluded: 6701 not mentioning all 6 TLX subscales
Papers that mentioned all 6 TLX subscales: 2769
Unobtainable: 23 requested from authors but not received 3 unable to find author contact details
Papers obtained online (2720) or from authors (23): 2743
Excluded: 1960 not values for all 6 subscales, not right publication types, or not in English 105 method not containing necessary info 38 duplicates or preliminary versions 29 had fewer than 5 participants 17 graphs too poor to read accurately
Papers that met the inclusion criteria: 594
Excluded: 34 reporting weighted TLX values only 20 using scales with fewer than 7 points
Papers reporting raw TLX values: 540

Figure 1. Paper selection.

[*Figure 1 Alt Text*: Successive exclusions of papers from the 9647 papers citing Hart and Staveland to the 540 included papers.]

Second, the 2769 resulting papers were looked up. While the vast majority of the papers could be accessed online in full text, 46 papers could not and were requested from the authors, 23 of whom supplied a full-text copy. In addition, author contact details could not be identified for three papers. That is, 26 (0.9%) of the 2769 papers were unobtainable.

Third, the inclusion criteria were matched against the content of the papers. A total of 2149 papers did not meet the inclusion criteria, most frequently because they did not report values for all six subscales (Figure 1). When reported, the TLX values were either tabulated or graphed. Seventeen papers were excluded because the graphs were too poor to enable accurate reading. Other papers were excluded because they did not provide methodological information such as the numerical endpoints of the rating scales.

Fourth, the initial plan was to review papers reporting raw TLX as well as papers reporting weighted TLX. However, it turned out that 560 papers reported raw TLX and 34 papers reported only weighted TLX. Due to the depreciation of the weighting process (Hendy et al., 1993; Nygren, 1991) and the minority of papers applying it, it was decided to focus exclusively on raw TLX. It also turned out that the papers used different scale formats for measuring TLX values. A 0-100 scale was the most common but 20 papers used scale formats with fewer than 7 response categories. These 20 papers were excluded because Preston and Colman (2000) found that scales with so few response categories tended to perform poorly, while scales with at least 6 response categories correlated better with one another.

3.3 Data analysis

The data analysis was conducted by the author and proceeded in four steps. First, the 540 papers were coded one by one. This involved extracting methodological information such as the number of participants, the numerical endpoints of the rating scales, and whether the TLX data were from a real-life or lab setting. It also involved extracting values for the TLX subscales. This was done for each condition for which such values were reported, with the exception that conditions with fewer than five participants were excluded. For each condition, the domain and technology were also qualitatively described. If a paper included multiple studies, they were coded separately. There were 14 papers with 2 studies and 1 paper with 3 studies, for a total of 556 studies.

Second, the descriptions of domains and technologies were classified into groups. This was done in a bottom-up process that resembled affinity diagramming (Beyer & Holtzblatt, 1998). For domains, the final grouping included 18 domains, see Table 1. If a study was in the overlap between two domains (e.g., military and aviation), it was grouped by its primary domain (e.g., military, if the study was about air combat; aviation, if the study was about instrumentation for safe landing). For technologies, the final grouping involved five groups that reflected the 'size' of the technology: handheld devices (e.g., handguns and mobile phones), desktop applications (e.g., websites and electronic health records), environments (e.g., cockpits and control rooms), virtual reality (in which participants wore glasses that immersed them in a virtual environment), and other.

Third, the values for the six subscales were rescaled to the 0-100 range and TLX was calculated as the mean of the subscales. Values were rescaled using the formula: (value - lower endpoint) / (max endpoint - lower endpoint) * 100. This formula corresponded to those used by Preston and Colman (2000) and Lewis and Erdinc (2017). It for example rescaled a 4 on a 1-7 scale into (4-1)/(7-1)*100 = 50.

Fourth, the TLX data were analyzed. In the analyses, each of the 556 studies contributed one MD, PD, TD, EF, PE, FR, and TLX value: the mean of the conditions reported for that study. This averaging served to make the analyses independent of how many conditions each study had. However, one analysis was made without the averaging. This analysis concerned the technologies, which often differed from one condition to another in the same study.

Table 1. The domain classification.

Domain	Description
Aviation	Control and operation of airplanes, including air traffic control
Driving	Control and operation of cars and other motor vehicles
Education	Studying at high school, university, and other learning institutions
Emergency response	Ambulance services, firefighting, police, and similar non-military services
Engineering design	Process of devising a system or process to meet desired needs
General	Activities not tied to a specific domain
GLAM	Galleries, libraries, archives, and museums
Healthcare	Hospital services provided by nurses, pharmacists, physicians, and others
Leisure	Non-work activities such as playing games, web surfing, and travelling
Manual labor	Physical work performed using basic implements rather than machines
Maritime	Control and operation of ships and other vessels
Military	Command and control, cyber operations, and other armed-forces activities
Office work	Clerical activities such as administrative tasks, data entry, and proofreading
Process control	Overseeing a power plant or similar facility from a control room

Production	Industrial manufacturing and assembly of products, including buildings
Space travel	Control and operation of spacecrafts
Special-needs users	Activities of people with cognitive, motor, visual, and other impairments
Sports	Competitive sports such as cycling, rugby, swimming, and table tennis

4 Results

The use of TLX for measuring workload increased over the 30-year period. Half of the included papers were published during the last four years of the period. The 556 studies reported TLX data from a total of 27616 participants.

4.1 Distribution of TLX and its subscales

The TLX values were in the range from 5 to 85, see Figure 2. This wide range occurred even though the value from each study was the mean across the participants in the study. The workloads experienced by individual study participants spanned an even wider range. The median TLX was 43 and the mean 42 (Table 2).

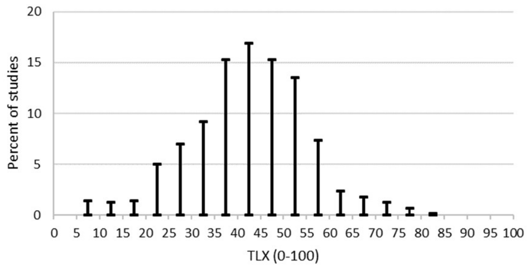

Figure 2. Distribution of TLX across 20 bins (0-5, 5-10, 10-15, etc.), *N* = 556 studies.

[*Figure 2 Alt Text*: Bell-shaped distribution of the TLX values around a median of 43.]

The graphs of the accumulated distribution of the subscales had the characteristic S-shape indicating a concentration of cases around the median, see Figure 3. FR had the highest concentration. The 50% of the studies in the interquartile range for FR spanned just 19 scale points. In contrast, the interquartile range was 27 for PE, which had the most even spread of values among the six subscales.

MD and EF had the highest medians, PD and FR the lowest. That is, the TLX values tended to be driven upward by MD and EF and downward by PD and FR (Figure 3). The two remaining subscales, TD and PE, drove TLX slightly downward below their median and slightly upward above their median. For all subscales, the median was close to the mean, see Table 2. The largest difference was three (for PE).

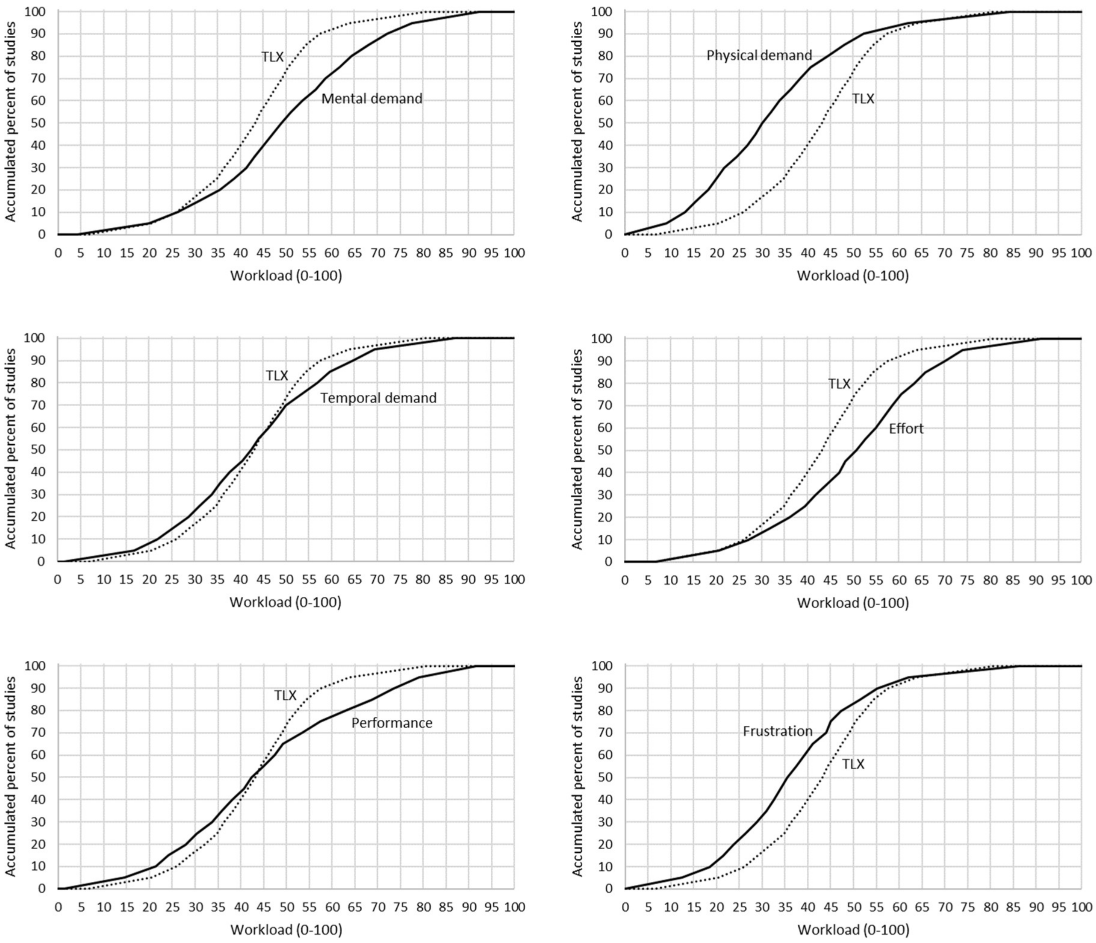

Figure 3. Distribution of the six TLX subscales (solid lines) and the TLX score (dotted line), *N* = 556 studies. The TLX score (the same in all six graphs) was included to show the contribution of each subscale to the overall score.

[*Figure 3 Alt Text*: Six graphs, each showing the S-shaped accumulated distribution of one subscale against that of the overall TLX score.]

Table 2. Percentiles, mean, and standard deviation (SD) of TLX and its subscales, *N* = 556 studies.

	MD	PD	TD	EF	PE	FR	TLX
10th percentile	26	13	22	27	21	18	26
20th percentile	36	18	29	36	28	24	32
30th percentile	41	22	34	42	34	29	36
40th percentile	45	27	38	47	38	32	40
50th percentile (median)	49	30	42	51	42	36	43
60th percentile	54	34	46	55	47	39	46
70th percentile	59	38	50	59	53	44	49
80th percentile	64	45	57	63	63	47	52
90th percentile	72	52	65	70	73	55	57
Mean ± SD	49±17	32±16	42±16	50±16	45±19	36±14	42±13

4.2 Domain variation

Unsurprisingly, the overall pattern in Figure 3 hid substantial variation across domains. Depending on the domain, TLX varied from 35 (leisure) to 56 (manual labor), see Table 3. A test of the 10 domains represented by at least 20 studies showed significant cross-domain variation in TLX, *F*(9, 477) = 3.56, *p* < .001. Similarly, MD, PD, TD, EF, PE, and FR varied significantly across domains, *F*(9, 477) = 3.89, 6.05, 4.42, 5.06, 2.23, and 3.23, respectively (all *p*s < .05).

The overall pattern that TLX tended to be driven upward by MD and EF and downward by PD and FR was refound for most domains (Table 3). MD was highest for two domains that also had high TLX (military and office work) and lowest for galleries, libraries, archives, and museums (GLAM) and leisure. Manual labor had by far the highest PD; it was three times that for several other domains. With a maximum of 42 (for process control), FR was the only subscale that did not exceed 60 for any domain. PE was best (i.e., lowest) for leisure.

Table 3. TLX values (mean ± standard deviation) for different domains, *N* = 556 studies.

Domain	Studies	MD	PD	TD	EF	PE	FR	TLX
Aviation	34	51±15	31±14	43±14	49±15	52±19	33±13	43±11
Driving	63	47±14	29±12	38±13	45±14	41±16	34±12	39±11
Education	30	55±13	34±13	45±11	58±10	52±17	38±11	47±7
Emergency response	12	52±17	33±23	49±13	54±13	49±18	37±7	46±11
Engineering design	18	47±17	23±14	35±15	43±17	53±26	30±13	38±12
General	127	48±17	30±13	43±15	50±15	42±16	40±14	42±12
GLAM	7	39±4	36±9	40±9	43±5	68±15	37±11	44±4
Healthcare	87	52±19	36±18	46±18	51±17	44±22	39±16	45±16
Leisure	21	41±15	23±14	37±13	39±14	36±17	32±14	35±12
Manual labor	2	44±27	67±4	61±8	69±5	56±34	38±11	56±12
Maritime	1	56	23	41	60	41	22	40
Military	26	62±11	25±12	51±12	60±9	43±12	41±11	47±8
Office work	8	58±22	33±12	51±17	53±17	51±21	37±12	47±15
Process control	20	56±21	27±15	47±22	50±19	49±21	42±20	45±17
Production	38	48±19	43±19	44±17	50±20	42±24	33±16	43±14
Space travel	4	50±8	23±13	44±14	53±6	39±21	29±6	39±6
Special-needs users	41	43±16	27±12	33±17	42±17	44±21	31±15	37±11
Sports	17	43±16	44±18	42±15	53±18	47±18	31±14	44±14

4.3 Variation across technologies

Handheld devices, desktop applications, environments, and virtual reality were investigated in 1024 conditions from 332 studies, see Table 4. A test of these technologies showed significant variation in TLX across technologies, *F*(3, 1020) = 12.08, *p* < .001. Bonferroni-adjusted pairwise comparisons showed that TLX was lower for handheld devices than for the three other technologies. This pattern was also present for MD and EF, *F*(3, 1020) = 30.44 and 19.46, respectively (both *p*s < .001).

The technologies were spread unevenly across domains. While desktop applications occurred in many domains, environments mainly occurred in driving (51%) and aviation (23%), handheld devices in general (51%) and leisure (12%), and virtual reality in education (28%) and healthcare (21%). Thus, variation in TLX across technologies was intermixed with variation across domains.

Table 4. TLX values (mean ± standard deviation) for different technologies, *N* = 1024 conditions.

Technology	Conditions	MD	PD	TD	EF	PE	FR	TLX
Handheld device	156	37±18	30±18	35±18	39±19	40±23	31±16	35±16
Desktop application	533	52±19	28±15	44±18	51±17	43±19	37±16	43±14
Environment	263	51±16	31±16	41±15	49±16	48±18	34±14	42±12
Virtual reality	72	47±20	26±13	40±17	48±19	47±24	34±18	41±15

4.4 Regional variation

TLX also varied with the region in which the studies were conducted, see Table 5. A test of the 3 regions represented by at least 20 studies showed significant regional variation in TLX, *F*(2, 529) = 13.69, *p* < .001. Bonferroni-adjusted pairwise comparisons showed higher TLX for studies in Asia than in Europe and North America. This pattern recurred within 3 of the 12 domains represented by studies from all 3 of these regions: healthcare, *F*(2, 82) = 14.60, *p* < .001, office work, *F*(2, 4) = 8.84, *p* < .05, and production, *F*(2, 34) = 8.21, *p* < .01.

The higher TLX for studies in Asia than in Europe and North America was not caused by any subscale in particular. Asia had higher values than Europe for MD, PD, TD, EF, and FR, *F*(1, 312) = 20.01, 33.64, 16.76, 19.75, and 11.72, respectively (all *p*s < .001). And higher values than North America for MD, PD, EF, and PE, *F*(1, 309) = 5.42, 49.30, 6.42, and 8.75, respectively (all *p*s < .05). Across all six regions, Asia and South America had the maximum subscale values, while Africa had the minimum values (Table 5). However, Africa and South America were only represented by few studies, thereby making the values less robust.

Table 5. TLX values (mean ± standard deviation) for regions, *N* = 556 studies.

Region	Studies	MD	PD	TD	EF	PE	FR	TLX
Africa	3	29±5	27±11	28±7	36±5	36±14	28±8	31±2
Asia	93	55±19	42±18	47±19	55±19	49±21	40±18	48±16
Australasia	16	49±16	30±16	42±15	50±17	39±15	37±15	41±14
Europe	221	46±16	31±15	39±15	46±15	45±20	34±13	40±11
North America	218	50±16	29±14	44±16	50±15	42±18	37±14	42±12
South America	5	65±9	35±20	48±12	58±11	64±12	40±9	52±9

4.5 Variation across settings

TLX was eight scale points higher when studied in real-life settings rather than lab settings, *F*(1, 554) = 27.21, *p* < .001 (Table 6). This difference recurred within three of the four domains represented by at least five studies for both real-life and lab settings: healthcare, *F*(1, 85) = 4.42, *p* < .05, process control, *F*(1, 18) = 12.26, *p* < .01, and production, *F*(1, 36) = 21.60, *p* < .001. Furthermore, the higher values for real-life settings were consistent across all subscales except PE, *F*(1, 554) = 23.77, 46.44, 28.46, 25.93, and 5.31 for MD, PD, TD, EF, and FR, respectively (all *p*s < .05).

Table 6. TLX values (mean ± standard deviation) for real-life and lab settings, *N* = 556 studies.

Setting	Studies	MD	PD	TD	EF	PE	FR	TLX
Real-life setting	79	58±18	42±20	51±19	58±17	46±25	40±17	49±16
Lab setting	477	48±16	30±14	41±15	48±16	44±18	36±14	41±12

For real-life settings, the subscales were moderately to strongly inter-correlated and all subscales correlated strongly with the TLX score, see Table 7. In addition, EF correlated more strongly with the three demand subscales than with PE and FR. For lab settings, the correlations among the subscales tended to be lower (Table 8). Specifically, PD and PE had lower correlations with all other subscales in lab than real-life settings. In the lab, these correlations ranged from .24 to .46 (PD) and .30 to .42 (PE). In real-life settings, the lowest subscale inter-correlation was the .43 correlation between PD and PE. The strongest subscale inter-correlation in both real-life and lab settings was between MD and EF.

Table 7. Correlations among TLX and subscales for real-life settings, *N* = 79 studies.

	MD	PD	TD	EF	PE	FR	TLX
Mental demand (MD)	-	.55	.79	.80	.49	.59	.84
Physical demand (PD)	.55	-	.62	.75	.43	.47	.77
Temporal demand (TD)	.79	.62	-	.75	.60	.72	.90
Effort (EF)	.80	.75	.75	-	.50	.61	.88
Performance (PE)	.49	.43	.60	.50	-	.59	.76
Frustration (FR)	.59	.47	.72	.61	.59	-	.80
Task load index (TLX)	.84	.77	.90	.88	.76	.80	-

Note: Pearson correlations (all *p*s < .001)

Table 8. Correlations among TLX and subscales for lab settings, *N* = 477 studies.

	MD	PD	TD	EF	PE	FR	TLX
Mental demand (MD)	-	.24	.68	.83	.38	.69	.83
Physical demand (PD)	.24	-	.43	.46	.30	.41	.60
Temporal demand (TD)	.68	.43	-	.75	.33	.67	.83
Effort (EF)	.83	.46	.75	-	.42	.69	.90
Performance (PE)	.38	.30	.33	.42	-	.32	.62
Frustration (FR)	.69	.41	.67	.69	.32	-	.81
Task load index (TLX)	.83	.60	.83	.90	.62	.81	-

Note: Pearson correlations (all *p*s < .001)

5 Discussion

Hancock and Matthews (2019, p. 388) have found that workload measurements are operationally useful because “insights from dimensions of workload beyond manifest performance itself can provide vital input”. The present study facilitates such insights by supplying reference values and subscale patterns for TLX across domains, technologies, regions, and settings.

5.1 Patterns in TLX

Five findings stand out from the analysis of the 556 studies. First, the TLX measurements were symmetrically distributed around a mean of 42. TLX tended to be driven upward by MD and EF and downward by PD and FR. This pattern was present in the dataset in general and for most domains.

The higher values for MD and lower values for PD indicate that TLX has mainly been used for measuring mental workload. Accordingly, many authors have referred to TLX as a measure of mental workload (e.g., Galy et al., 2018; Young et al., 2015) and some have even dropped the PD subscale (e.g., Grubb et al., 1995; Haller et al., 2011). FR was the subscale with the lowest mean; it did not exceed 42 for any domain. The modest mean values for FR are encouraging from a technology usability perspective and provide some contrast to the finding by Lazar et al. (2006) that computer users wasted 42% of their time on computers due to frustrating experiences.

Second, TLX varied substantially across domains. This finding accords with Grier (2015), who cautiously noted, but did not study, that workload is influenced by factors beyond the domain. However, domain variation is important because most practitioner interest in workload is domain specific. Similarly, job redesign, technology diffusion, staff training, and other interventions that may influence workload normally occur within domains. For all six subscales, the variation across domains was at least as large as the domain variation in overall TLX.

Third, TLX varied across technologies in that handheld devices were associated with lower workload than other technologies. A candidate reason for this difference is that handheld devices are simpler because they are more task-specific. However, technology-imposed workload variation is difficult to disentangle from domain-imposed variation because desktop applications were the only technology studied across a wide range of domains. One way of investigating technology-imposed variation further is to restrict the investigation to a selected domain. Such an investigation could review the many within-domain comparisons of the workload imposed by different technologies (e.g., Chao et al., 2017; Hertzum & Simonsen, 2016; Yan et al., 2017). Narrower categories of technology should also be considered because they may reveal additional variation in workload.

Fourth, TLX was 6-8 scale points higher for studies in Asia than in Europe and North America. This pattern recurred within healthcare, office work, and production and, thus, partly reflected regional variation above and beyond the domain variation. In a comparison of 82 Dutch and 84 Indonesian students on memory search tasks, Johnson and Widyanti (2011) found that a 10-point difference in TLX (Dutch: 53, Indonesian: 63) merely approached significance. The present study suggests that the difference was real but masked by insufficient sample size. Therefore, regional variation in TLX should be considered in cross-country workload comparisons. Possible reasons for such variation include socio-cognitive differences in how people with different cultural backgrounds think (Nisbett, 2003). However, regional variation in workload warrants further study. Specifically, this study highlights the uneven spread of TLX data across regions. Studies should also control for performance to investigate whether regional differences in, for example, work ethic (i.e., working more or less diligently) drive both workload and performance. Johnson and Widyanti (2011) found no difference in performance.

Fifth, TLX was higher when studied in real-life as opposed to lab settings. This finding extends Niforatos et al. (2018), who studied experience-enhancing skiing helmets. The present study found higher workload in real-life settings overall and within healthcare, process control, and production. For most of the other domains, real-life settings have been studied too little to make domain-specific comparisons between settings. Possible explanations for the higher TLX in real-life settings include more multitasking and genuine consequences. In addition, the subscale correlations tended to be lower in lab settings, particularly the correlations involving PD or PE. That is, PD and PE were somewhat dissociated from the other subscales in lab settings compared to real-life settings.

5.2 Using the reference values

Practitioners can use the reference values supplied by this study in three ways. First, measurements of TLX and its subscales can be compared against Table 2 to get an overall sense of how high or low the measurements are. To account for the domain, technology, and region, the measurements can be compared against Tables 3-5, which can also be used to adjust the reference values in Table 2 upward or downward. Second, the contributions of the individual subscales to the overall TLX score can be compared with the subscale patterns in the reference values. While Hart and Staveland (1988) documented that the subscales capture somewhat independent dimensions of workload, the reference values indicate patterns that are frequent across the 556 studies. Third, Table 6 suggests that TLX measurements obtained in the lab should be adjusted upward to reflect the workload in real-life settings. Relatedly, lab measurements may underestimate the subscale inter-correlations (Tables 7 and 8).

While TLX is mostly measured on a 0-100 scale, practitioners also use other scale formats. A few papers provide a rationale for their choice of scale format, for example that a format with few scale points is preferable because the participants are special-needs users (e.g., Funk et al., 2015). These papers apart, the present study recommends consistent use of the 0-100 scale format. If another scale format is used, the reference values in this paper can be rescaled using the formula: reference value / 100 * (max endpoint - lower endpoint) + lower endpoint. This formula, for example, rescales 42 to 42/100*(7-1) + 1 = 3.52 on a 1-7 scale.

5.3 Limitations

This study has three limitations that should be kept in mind. First, the distribution of the 556 studies across domains, technologies, regions, and settings is not balanced. Consequently, variation across, say, technologies may fully or partly reflect underlying co-variation between technology and another variable, such as domain. Possible co-variations have been investigated and reported for all analyses, but in the absence of a balanced design they cannot be systematically partialed out. Second, the domain and technology classifications group studies together if they share specific domain and technology characteristics, respectively. However, some of the studies that are grouped together are more similar than others. For example, handguns and mobile phones are both classified as handheld devices but it may be questioned whether the workload involved in firing a handgun resembles that of using a mobile phone. Third, variables other than domain, technology, region, and setting may influence TLX measurements and cause the variation across domains, technologies, regions, and settings. For example, the translation of the TLX instrument may cause regional variation. Differences in tasks between lab and real-life studies may cause variation across settings. A review cannot control such variables fully. Rather, this review points to region and setting as two variables that warrant further attention. Future studies should investigate the mechanisms that underlie variation in TLX across regions and settings.

6 Conclusion

This study provides reference values for TLX and its subscales. These reference values account for variation in TLX across domains, technologies, regions, and real-life/lab settings. Researchers and practitioners can use the reference values to benchmark their own TLX measurements against those from the corpus of TLX research. The study also investigates subscale patterns, which expand on TLX values and, thereby, add to the diagnostic power of the instrument. In summary, it is hoped that the reference values and subscale patterns will assist in interpreting TLX measurements and motivate future work on the mechanisms that underlie variation in TLX measurements.

Disclosure statement

The author has no conflicts of interest to declare.

References

Beyer, H., & Holtzblatt, K. (1998). *Contextual design: Defining customer-centered systems*. San Francisco, CA: Morgan Kaufmann.

Bliss, J.P., & Hanson, J.A. (2018). The effects of task criticality and target modality on a simulated battlefield search task. *Military Psychology, 30*(2), 108-119. https://doi.org/10.1080/08995605.2017.1420981

Chao, C.-J., Wu, S.-Y., Yau, Y.-J., Feng, W.-Y., & Tseng, F.-Y. (2017). Effects of three-dimensional virtual reality and traditional training methods on mental workload and training performance. *Human Factors and Ergonomics in Manufacturing & Service Industries, 27*(4), 187-196. https://doi.org/10.1002/hfm.20702

de Waard, D., & Lewis-Evans, B. (2014). Self-report scales alone cannot capture mental workload. *Cognition, Technology & Work, 16*(3), 303-305. https://doi.org/10.1007/s10111-014-0277-z

de Winter, J.C.F. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology & Work, 16*(3), 289-297. https://doi.org/10.1007/s10111-014-0275-1

Epps, J. (2018). Task load and stress. In K.L. Norman & J. Kirakowski (Eds.), *The Wiley Handbook of Human Computer Interaction. Volume 1* (pp. 207-223). Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118976005.ch11

Funk, M., Bächler, A., Bächler, L., Korn, O., Krieger, C., Heidenreich, T., & Schmidt, A. (2015). Comparing projected in-situ feedback at the manual assembly workplace with impaired workers. In *Proceedings of the PETRA2015 Conference on Pervasive Technologies Related to Assistive Environments* (paper 1). New York: ACM Press. https://doi.org/10.1145/2769493.2769496

Galy, E., Paxion, J., & Berthelon, C. (2018). Measuring mental workload with the NASA-TLX needs to examine each dimension rather than relying on the global score: An example with driving. *Ergonomics, 61*(4), 517-527. https://doi.org/10.1080/00140139.2017.1369583

Grier, R.A. (2015). How high is high? A meta-analysis of NASA-TLX global workload scores. In *Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting* (pp. 1727-1731). Santa Monica, CA: HFES. https://doi.org/10.1177/1541931215591373

Grubb, P.L., Warm, J.S., Dember, W.N., & Berch, D.B. (1995). Effects of multiple-signal discrimination on vigilance performance and perceived workload. In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting* (pp. 1360-1364). Santa Monica, CA: HFES. https://doi.org/10.1177/154193129503902101

Haller, M., Richter, C., Brandl, P., Gross, S., Schossleitner, G., Schrempf, A., Nii, H., Sugimoto, M., & Inami, M. (2011). Finding the right way for interrupting people: Improving their sitting posture. In *Proceedings of the INTERACT2011 Conference on Human-Computer Interaction* (Vol. LNCS 6947, pp. 1-17). Berlin: Springer. https://doi.org/10.1007/978-3-642-23771-3_1

Hancock, P.A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human Factors, 61*(3), 374-392. https://doi.org/10.1177/0018720818809590

Hart, S.G. (2006). NASA-task load index (NASA-TLX): 20 years later. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 904-908). Santa Monica, CA: HFES. https://doi.org/10.1177/154193120605000909

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139-183). Amsterdam: North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hendy, K., Hamilton, K.M., & Landry, L.N. (1993). Measuring subjective workload: When is one scale better than many? *Human Factors, 35*(4), 579-601. https://doi.org/10.1177/001872089303500401

Hertzum, M., & Simonsen, J. (2016). Effects of electronic emergency-department whiteboards on clinicians' time distribution and mental workload. *Health Informatics Journal, 22*(1), 3-20. https://doi.org/10.1177/1460458214529678

Johnson, A., & Widyanti, A. (2011). Cultural influences on the measurement of subjective mental workload. *Ergonomics, 54*(6), 509-518. https://doi.org/10.1080/00140139.2011.570459

Krekhov, A., & Krüger, J. (2019). Deadeye: A novel preattentive visualization technique based on dichoptic presentation. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 936-945. https://doi.org/10.1109/TVCG.2018.2864498

Lazar, J., Jones, A., & Shneiderman, B. (2006). Workplace user frustration with computers: An exploratory investigation of the causes and severity. *Behaviour & Information Technology, 25*(3), 239-251. https://doi.org/10.1080/01449290500196963

Lewis, J.R., & Erdinc, O. (2017). User experience rating scales with 7, 11, or 101 points: Does it matter? *Journal of Usability Studies, 12*(2), 73-91.

Niforatos, E., Fedosov, A., Langheinrich, M., & Elhart, I. (2018). Augmenting humans on the slope: Two electronic devices that enhance safety and decision making. *IEEE Consumer Electronics Magazine, 7*(3), 81-89. https://doi.org/10.1109/MCE.2018.2797718

Nisbett, R.E. (2003). *The geography of thought: How Asians and Westeners think differently - and why*. London: Brealey.

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. https://doi.org/10.1177/001872089103300102

Preston, C.C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1-15. https://doi.org/10.1016/S0001-6918(99)00050-5

Reid, G.B., & Nygren, T.E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 185-218). Amsterdam: North-Holland. https://doi.org/10.1016/S0166-4115(08)62387-0

Tsang, P.S., & Velazquez, V.L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics, 39*(3), 358-381. https://doi.org/10.1080/00140139608964470

Vidulich, M.A., & Tsang, P.S. (2012). Mental workload and situation awareness. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics. Fourth Edition* (pp. 243-273). Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118131350.ch8

Wierwille, W.W., & Casali, J.G. (1983). A validated rating scale for global mental workload measurement application. In *Proceedings of the Human Factors Society 27th Annual Meeting* (pp. 129-133). Santa Monica, CA: HFS. https://doi.org/10.1177/154193128302700203

Yan, S., Tran, C.C., Chen, Y., Tan, K., & Habiyaremye, J.L. (2017). Effect of user interface layout on the operators' mental workload in emergency operating procedures in nuclear power plants. *Nuclear Engineering and Design, 322*, 266-276. https://doi.org/10.1016/j.nucengdes.2017.07.012

Young, M.S., Brookhuis, K.A., Wickens, C.D., & Hancock, P.A. (2015). State of science: Mental workload in ergonomics. *Ergonomics, 58*(1), 1-17. https://doi.org/10.1080/00140139.2014.956151