Behaviour & Information Technology, Vol. 41(2022, No. 16, pp. 3506-3518.
Author version. Published version at: https://doi.org/10.1080/0144929X.2021.2000642

Associations among Workload Dimensions, Performance, and Situational Characteristics: A Meta-Analytic Review of the Task Load Index

Morten Hertzum

University of Copenhagen, Karen Blixens Plads 8, Copenhagen, Denmark, hertzum@hum.ku.dk

ORCID: 0000-0003-0019-8531

Abstract. Workload is an important explanatory variable in human-computer interaction and commonly measured with the Task Load Index (TLX). Thus, it is important to understand the qualities of TLX and its relations to other variables. By reviewing 384 papers that apply TLX, this study analyzes how differences in TLX and its six subscales are associated with one another and with differences in performance, user experience, and situational characteristics. Six findings stand out. First, the TLX subscales measure associated, but somewhat independent, dimensions of workload. Second, people compensate for demanding conditions by putting in more effort and, as a result, sometimes avoid a drop in performance. Third, differences in workload are associated with differences in error rate, task completion time, and user experience but the strength of association is merely slight to fair. Fourth, differences in opposite directions between workload and either error rate, task completion time, or user experience are few but occur for all TLX subscales. Fifth, differences in workload dimensions are more often associated with differences in tasks and contexts than users and systems. Sixth, the TLX subscales – not just the composite TLX score – are widely used for testing cross-system differences in workload.

Keywords: mental workload, NASA-TLX, Task Load Index, TLX, workload

1 Introduction

Workload is a frequent explanatory variable in studies of human-computer interactions (Epps, 2018). It helps explain task performance (e.g., Melman et al., 2017; Rasmussen & Hertzum, 2013) as well as user experience (e.g., Katsuragawa et al., 2017; Loup-Escande et al., 2017). Workload 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). That is, it is about the balance, or imbalance, between the externally imposed requirements and the user's capabilities. The predominant instrument for measuring workload is the Task Load Index (TLX, aka NASA-TLX; Hart & Staveland, 1988).

TLX measures self-reported workload. Other classes of workload measurement include analytic techniques such as expert assessment, empirical measures such as secondary task performance, and physiological indicators such as heart rate (Gawron, 2019). However, self-report measures and, in particular, TLX have become so widely used that de Winter (2014, p. 293) stated that "workload has become synonymous with the TLX". One reason for the widespread use of self-report measures is that experienced workload has genuine consequences: People who experience their workload as excessive will behave as though they are overloaded, even if the imposed requirements are objectively low (Hart & Staveland, 1988). The widespread use of TLX makes it possible to follow up on its psychometric properties by reviewing published papers that apply the TLX instrument.

The present study reviews 384 papers to analyze how differences in TLX and its subscales are associated with one another, with differences in performance, and with differences in situational characteristics. The TLX subscales are important because they represent different workload dimensions and are believed to add diagnostic power to the workload measurement (Galy et al., 2018; Hart & Staveland, 1988). Many papers in human-computer interaction focus on the subscales to the extent of not even reporting the composite TLX score (e.g., Majrashi, 2019; Prilla et al., 2019; Son & Lee, 2017). These papers were included in this review; papers that merely reported the composite TLX score were not. Specifically, this study aims to answer three research questions:

1. To what extent do the TLX subscales change together as opposed to separately?
2. To what extent are differences in TLX values in agreement with differences in error rate, task completion time, and user experience?
3. To what extent are differences in TLX values associated with differences in the characteristics of users, tasks, systems, and contexts?

The rationale for the first research question is to look into the diagnostic power of the TLX subscales. This question concerns whether the subscale values provide additional information and should be reported along with the composite TLX score, as recommended by Galy et al. (2018). The second research question is about how often TLX helps explain differences in the performance and experience of tasks. This question is important because TLX is often employed for this purpose and because the relation between workload and, in particular, performance has been debated (Hancock & Matthews, 2019). The third research question is about how to change workload. The answer to this question is practically important because practitioners would benefit from knowing whether changes in the characteristics of users, tasks, systems, or contexts are more likely to influence workload.

2 The TLX instrument

Workload is an influential and debated concept (de Winter, 2014; Dekker & Hollnagel, 2004; Hancock & Matthews, 2019; Parasuraman et al., 2008; Young et al., 2015). The debate concerns whether workload is too slippery to constitute a robust scientific concept or sufficiently well-defined to provide operationally useful insights. The measurement of workload is central to the debate because measurements co-define the concept. With its six subscales, TLX makes workload a multidimensional concept. While most other self-report measures of workload are also multidimensional, their dimensions vary (Gawron, 2019). The six TLX subscales are mental demand, physical demand, temporal demand, effort, performance, and frustration, see Table 1. In developing TLX, Hart and Staveland (1988) validated that the subscales measure somewhat independent dimensions of workload and that TLX provides “a sensitive and reliable estimate of workload” (p. 139).

Each subscale is measured with a single item. Hence, the entire instrument consists of six items and is, therefore, easy to administer. The items are rated by marking a response between the endpoints ‘Low’ and ‘High’, with the exception that the performance subscale has the endpoints ‘Good’ and ‘Poor’ (note that a numerically higher performance rating indicates poorer performance). Often, the composite TLX score is simply the average of the six item ratings. This way of calculating the composite score is known as raw TLX (Hart, 2006). To help interpret TLX values, Hertzum (2021) provides reference values for the subscales and for raw TLX.

The alternative to raw TLX is weighted TLX, which tailors the TLX instrument to the task by assigning more weight to the most important dimensions of the task. The weighting is made after the rating of the subscales and consists of indicating the more important subscale in each of the 15 possible pairs of subscales. On this basis, weighted TLX is calculated in two steps. First, the weighted contribution of each subscale is its rating times its weight (the weight is the number of pairs in which the subscale is deemed the more important). Second, weighted TLX is the sum of the weighted subscale contributions divided by 15. The weighting procedure makes TLX more complex to administer. Some studies find that the weighting procedure is ineffective and therefore unnecessary (Byers et al., 1989; Nygren, 1991).

3 Method

Following procedures for systematic reviews, the author selected and analyzed 384 papers, which contained 904 tests of TLX subscales across pairs of study conditions.

3.1 Paper selection

Seven inclusion criteria guided the paper-selection process. To be included in the analysis, a paper had to satisfy all seven criteria:

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

The first criterion specified that the papers had to report statistical tests of the TLX subscales. Papers were excluded if they merely reported a test of the composite TLX score or if they merely reported an omnibus test without also reporting pairwise comparisons. Tests of pairs of study conditions were necessary for full transparency about the situational characteristics of the compared conditions. The second criterion ensured that all included papers defined TLX in the same way. The third, fourth, and fifth criteria served to bolster the quality of the data set. The two last criteria set the limits of the data set.

The inclusion criteria were applied in a four-step process, see Figure 1. First, Google Scholar was searched for the papers that cited Hart and Staveland (1988), were published in the 1990-2019 period, and contained the terms for the six subscales (i.e., “mental demand”, “physical demand”, “temporal demand”, “effort”, “performance”, and “frustration”).

Second, the full text of the 2769 resulting papers was looked up online. For 46 papers, the full text could not be accessed online. These papers were requested from the authors, 23 of whom supplied a full-text copy. For an additional three papers, author contact details could not be identified. That is, 26 (0.9%) of the 2769 papers were unobtainable.

Third, the content of the papers was matched against the inclusion criteria. As a result, 2339 papers were excluded, most frequently because they did not test the TLX subscales across pairs of study conditions (Figure 1). To avoid mixing up papers based on raw and weighted TLX scores, the 34 papers that tested weighted subscales values across conditions were also excluded. The low number of such papers probably reflected the depreciation of the weighting process (Byers et al., 1989; Nygren, 1991). Possibly, the weighting process is more common in papers that focus on the composite TLX score rather than report tests for the individual subscales.

Fourth, it turned out that the papers used different scale formats to measure TLX values. While a 0-100 scale was the most common, 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 compared to scales with at least 6 response categories.

3.2 Data analysis

The data analysis proceeded as follows. First, methodological information was extracted from each of the 384 included papers. This involved information such as the number of participants, the country in which the study was conducted, and the scale format of the rating scales. The remainder of the coding involved extracting information about the tests that compared the workload in one study condition with that in another study condition. Many papers reported from studies with more than two conditions and, thus, contained multiple pairwise comparisons. In total, the 384 papers contained 904 tests of the TLX subscales across pairs of study conditions.

Second, the outcome of each of the 904 tests was coded for the six subscales and, if present, for the composite TLX score. The coding distinguished among three test outcomes: (1) the test did not find a significant difference between the two study conditions, (2) the test found a significant difference in favor of the first condition (i.e., lower mental demand, lower physical demand, etc.), and (3) the test found a significant difference in favor of the second condition.

Third, the associations among the subscales were calculated on the basis of this coding of the test outcome. For each pair of subscales the coding gave a 3x3 cross-tabulation of the frequencies of the outcome combinations. As measures of association, we used the percentage agreement (i.e., the proportion of tests in the diagonal of the cross-tabulation) and Cohen's (1960) Kappa. A Kappa value above zero indicates statistically significant agreement. In addition, a higher Kappa value indicates stronger agreement. Landis and Koch (1977) proposed labels for the strength of agreement expressed by different ranges of Kappa values: "slight" (.00 - .20), "fair" (.21 - .40), "moderate" (.41 - .60), "substantial" (.61 - .80), and "almost perfect" (.81 - 1.00).

Fourth, the three same test-outcome categories were used to code whether the study conditions differed in terms of error rate, task completion time, and user experience, if tests for these variables were present. This coding was used to analyze the extent to which differences in workload were associated with differences in error rate, task completion time, and user experience. Error rate was defined as the proportion of tasks that were unsuccessful because the result was wrong or the task not completed. Task completion time was the time spent solving a task. User experience was the participants' perception of how it had been to solve the task, measured using instruments such as AttrakDiff (Hassenzahl et al., 2003), the System Usability Scale (Brooke, 1996), and various study-specific rating scales.

Fifth, the independent variable in each of the 904 tests was coded. This coding distinguished among four categories (Table 2): user, task, system, and context. In 16 tests, the independent variable involved two of these categories. For example, Smith et al. (1995) compared the workload experienced by reactor operators and maintenance operators at a nuclear power plant when these two user groups were performing their role-specific tasks. Thus, the independent variable involved both user and task. The other 888 tests were coded with one of the categories. With this coding, it was possible to analyze whether differences in workload were more often associated with differences in some of the four situational characteristics than in others.

4 Results

The 384 papers included a total of 18862 participants and contained 904 tests that compared the workload in a pair of study conditions. Geographically, these pairwise tests were distributed across Europe (367), North America (341), Asia (161), Australasia (32), South America (2), and Africa (1).

4.1 Associations among subscales

Depending on subscale, between 25% and 36% of the tests were significant (Table 3). The remaining tests – the majority – compared conditions that were too similar to yield a significant difference in workload. The percentage of significant tests varied across the subscales, $F(5, 889) = 10.66$, $p < .001$.

Bonferroni-adjusted pairwise comparisons showed that tests of mental demand, physical demand, effort, and frustration were more often significant than tests of performance and that tests of mental demand and effort were more often significant than tests of temporal demand.

Table 4 shows the associations among TLX and its subscales. All associations were statistically significant (Kappa > 0), but the strength of agreement was merely fair to moderate (.24 - .61). The agreement percentages ranged from 65% to 78%. Inspection of the data showed that most disagreements were tests that differed significantly on one of the subscales, but did not differ significantly on the other. However, there were instances of significant differences in opposite directions for all pairs of subscales. For example, rugby athletes experienced higher mental demand but lower frustration when they received real-time performance feedback during physical training exercises than when they received no feedback (Wilson et al., 2017). And participants experienced higher temporal demand but lower physical demand when using a semi-autonomous drive-safe system rather than a cane for navigation assistance during wheelchair driving (Sharma et al., 2012).

4.2 Association with error rate

A subset of the 904 tests that compared the workload of two study conditions also compared the error rate of the two study conditions. For these tests, the association between workload and error rate could be assessed. Table 5 shows the results. Appendix A provides additional detail.

There was a statistically significant association between workload and error rate for TLX and all its subscales (Table 5, agreement across all tests). However, the strength of agreement was merely slight to fair (.16 - .32), as also indicated by agreement percentages in the 58-65% range. Most disagreements were tests that differed significantly for either workload or error rate, but not both (Appendix A). For example, participants performed at lower error rates but experienced no difference in composite TLX score when they used a virtual-reality system rather than a technical manual for training vehicle-maintenance tasks (Chao et al., 2017).

To analyze the prevalence of significant differences in opposite directions, we also calculated agreement percentages and Kappa values by looking only at the tests that found a significant difference in both workload and error rate (Table 5, agreement across significant tests). For these tests, the strength of agreement between workload and error rate was substantial to almost perfect (.67 - .87), except for temporal demand (.56). Hence, the vast majority of the significant differences were in the same direction. However, there were instances of significant differences in opposite directions for all subscales. The number of such instances ranged from 5% (frustration) to 16% (temporal demand). For example, drivers experienced higher temporal demand but made fewer driving errors when they played a tailor-made game while driving than when they drove without playing (Bier et al., 2019). The game consisted of challenges related to the real-time driving conditions, such as maintaining the minimum legal distance to the vehicle in front as accurately as possible. To avoid disrupting the driver’s focus on the road, the game instructions were projected on the windscreen.

4.3 Association with task completion time

Table 6 shows statistically significant agreement between workload and task completion time for TLX and all its subscales. The pattern was similar to that for error rate. The strength of agreement across all tests of workload and task completion time was slight to fair (.16 - .30) with agreement percentages in the 51-58% range. Most disagreements were tests that differed significantly for either workload or task completion time, but not both (Appendix A). For example, participants experienced higher mental, physical, and temporal demands when entering text on a handheld device while walking as opposed to sitting, but there was no difference in their text-entry speed (Dai et al., 2009).

For the tests that found a significant difference in both workload and task completion time, the strength of agreement was substantial (.63 - .78) with the only exception that mental demand just reached almost perfect agreement (.81). Like for error rate, there were instances of significant differences in opposite directions for all subscales. The number of such instances ranged from 9% (mental demand) to 18% (temporal demand). For example, screen-reader users experienced lower mental demand when they identified images on the basis of non-speech sounds (audemes) than conventional alt texts, but their task completion times were longer with the audemes (Thapa et al., 2017). Like alt texts, the audemes were developed for the individual image, and their length matched the duration of the synthetic reading of the alt texts. As an example, the audeme for the image of a river was the sound of flowing water.

4.4 Association with user experience

Table 7 shows statistically significant agreement between workload and user experience for TLX and all its subscales. The strength of agreement across all tests of workload and user experience was fair, except that temporal demand (.19) and TLX (.41) fell just outside this range. The agreement percentages were in the 56-67% range. Like for error rate and task completion time, most disagreements were tests that differed significantly for either workload or user experience, but not both (Appendix A).

For the tests that found a significant difference in both workload and user experience, the strength of agreement was moderate for performance (.50) and TLX (.60) and substantial for all other subscales (.64 - .75). Like for error rate and task completion time, there were instances of significant differences in opposite directions for all subscales. The number of such instances ranged from 12% (effort) to 25% (performance) and, thus, tended to be more common than for error rate and task completion time. For example, participants who walked on a treadmill during gait rehabilitation rated their performance as poorer but their user experience as more attractive, novel, and stimulating when the walking was gamified through a virtual-reality enhancement than when they walked on the treadmill without the virtual-reality enhancement (Kern et al., 2019).

4.5 Association with situational characteristics

The coding of the independent variable divided the 904 tests into those that compared different user groups, different tasks, different systems, and different contexts. In the following, we excluded the 16 tests where the independent variable involved more than one of these situational characteristics, leaving 888 tests for analysis. Depending on subscale and situational characteristic, the proportion of tests that found a significant difference in workload was between 20% and 55% (Table 8). The proportion differed significantly across situational characteristics, and it did so for TLX and all its subscales: mental demand, $F(3, 882) = 9.88$, $p < .001$, physical demand, $F(3, 879) = 3.04$, $p < .05$, temporal demand, $F(3, 882) = 11.36$, $p < .001$, effort, $F(3, 884) = 2.88$, $p < .05$, performance, $F(3, 882) = 3.03$, $p < .05$, frustration, $F(3, 882) = 2.93$, $p < .05$, and TLX, $F(3, 447) = 3.16$, $p < .05$. Bonferroni-adjusted pairwise comparisons showed that:

- For mental demand, tests comparing tasks were more often significant than tests comparing users
- For mental and temporal demand, tests comparing tasks were more often significant than tests comparing systems
- For mental demand and TLX, tests comparing contexts were more often significant than tests comparing users
- For physical demand, temporal demand, and performance, tests comparing contexts were more often significant than tests comparing systems

In short, differences in workload were more often associated with differences in tasks (e.g., Abich et al., 2017) and contexts (e.g., Jazani et al., 2016) than in users (e.g., Gao et al., 2018) and systems (e.g., Yu & Liu, 2010).

5 Discussion

Hart and Staveland (1988) developed and validated TLX in a series of 16 studies that involved a total of 247 participants. The present review extends the validation of TLX by including 384 studies with a total of 18862 participants. In the following, we first discuss the findings and then how TLX defines workload.

5.1 Findings

The results of this study can be summarized in six findings. The first two findings concern the first research question, the next two findings the second research question, and the last two findings the third research question.

First, *the six TLX subscales measure associated, but somewhat independent, dimensions of workload*. The review establishes this finding by analyzing tests that compare the workload in pairs of study conditions. In these tests, the test outcome for one subscale is significantly associated (Kappa > 0) with the test outcome for any other subscale but the strength of agreement is no higher than moderate (Kappa ≤ .61). It is a well-accepted threshold in the interpretation of Kappa that two measures are not similar to the extent of measuring the same construct if their Kappa value is below .60 (Lazar et al., 2017). The associations among mental demand, effort, and the composite TLX score border on this threshold (Table 4), thereby suggesting that mental demand and effort are similar constructs with a meaning close to that of the composite TLX score. The four other subscales have test outcomes that are more dissimilar from one another and from the composite TLX score. That is, these subscales are to a larger extent independent workload dimensions. This finding corroborates Hart and Staveland (1988, Table 11), who find higher correlations among mental demand, effort, and TLX than between any other pair of subscales. It further corroborates the associations among the subscales that differences in opposite directions are rare, though they occur for all pairs of subscales.

Second, *people compensate for more demanding conditions by putting in more effort and, as a result, sometimes avoid a drop in performance*. The basis for this finding is the significantly lower incidence of change in the performance subscale than in mental demand, physical demand, effort, and frustration. Behavioral compensation to maintain performance by putting in more effort shows that workload mediates between task demands and task performance. This mediating role is well known and an important reason for the interest in workload (Hancock & Warm, 1989; Hockey, 1997; Young et al., 2015). It indicates that the performance subscale, in particular, is somewhat independent from the other subscales. In safety-critical and competitive environments, it may be essential to avoid a drop in performance. A user's workload indicates whether the user is strained and at risk of being unable to compensate or at a level of workload where compensation is possible without increasing the risk of errors. The former often leads to the adoption of suboptimal coping strategies (Barnes & Van Dyne, 2009), the latter allows for unanticipated events that temporarily increase workload (Hancock & Warm, 1989).

Third, *differences in workload are associated with differences in error rate, task completion time, and user experience but the strength of association is merely slight to fair*. This finding reiterates that workload mediates between task demands and task performance; it does not mirror error rate, task completion time, and user experience. Depending on subscale, 58-65%, 51-58%, and 56-67% of the tests for a difference in workload agree with the test for a difference in error rate, task completion time, and user experience, respectively. Specifically, the performance subscale does not stand out from the other subscales, except that the association between differences in error rate and the performance subscale is marginally stronger than that between differences in error rate and the other subscales (Table 5). That is, the performance subscale cannot substitute for measurements of error rate and task completion time. Rather, it – like the other subscales – helps explain error rates and task completion times (e.g., Blane et al., 2018). Relatedly, the composite TLX score cannot substitute for user-experience measurements, even though the experienced workload is part of the user experience.

Workload contributes to the pragmatic component of user experience but its hedonic component adds qualities beyond workload (Hassenzahl, 2004; van Schaik & Ling, 2008).

Fourth, *differences in opposite directions between, on the one hand, workload and, on the other hand, error rate, task completion time, or user experience are few but occur for all TLX subscales*. The differences in opposite directions occur in studies of both overload (e.g., Matthews & Campbell, 2009) and underload (e.g., Bier et al., 2019). In relation to overload, Matthews and Campbell (2009) explain the higher error rate but lower effort rating during the tasks that exposed the users to excessive time pressure as a coping strategy. The users tended to avoid the excessive time pressure by reducing their effort and task engagement, thereby accepting a higher error rate. In relation to underload, Bier et al. (2019) explain the higher temporal-demand ratings but fewer driving errors when the users played a tailor-made game while driving as an increase in the users' attention to their driving. Because the game was about the users' real-time driving conditions, it engaged them in their driving and, thereby, led them to experience higher temporal demands and perform fewer errors. Underload can be just as detrimental to performance as overload (Young et al., 2015) but fits less neatly into the demand/capability model of workload because this model does not explain why an excess in capability should result in poor performance. Bier et al. (2019) suggest that the explanation involves a link from low workload through monotony to reduced task engagement.

Fifth, *differences in workload dimensions are more often associated with differences in tasks and contexts than in users and systems*. This finding suggests that changing the demand side in the demands/capability model is more effective at influencing workload than changing the capability side, with systems occupying a mediating role between the two sides. Possibly, the demands imposed by tasks and contexts can vary over a wider range than the capabilities held by users. For example, only two of twelve tests that compare composite TLX scores for more and less experienced users find that workload differs across experience levels (Tran et al., 2018; Wang et al., 2010). Alternatively, TLX may be more sensitive to changes in demands than capabilities. This alternative explanation is adopted by McKendrick and Cherry (2018, p. 44), who contend that TLX measures "perceived task difficulty." With significant workload differences in 20% (mental demand) to 31% (physical demand) of the tests across user characteristics, the present review shows that TLX is sensitive to user capabilities and, thus, not restricted to perceived task difficulty. However, significant differences in workload are more common in tests across tasks (28-49%), contexts (32-55%), and to some extent systems (20-41%).

Sixth, *the TLX subscales are widely used for testing cross-system differences in workload*. As much as 487 (54%) of the 904 tests compare workload across systems. Systems mediate between task demands and user capabilities by changing the support that users receive in performing tasks. These changes are associated with differences in all TLX subscales. For example, 28% of the cross-system tests find a difference in frustration, which is an important concept on its own (Bessière et al., 2006) and influences usability (e.g., Kratz et al., 2010) and user experience (e.g., Partala & Salminen, 2012) in addition to workload. In 255 (52%) of the 487 cross-system tests, the comparison involves the subscales only. These cases show that in human-computer interaction the TLX instrument is often used as an inventory of individual workload dimensions, rather than as an instrument for obtaining a composite workload score. Some authors even redefine or leave out some subscales to tailor TLX to their study; Hart (2006) recommends that such revised versions of the instrument should not be referred to as TLX. They are excluded from the present review.

5.2 Workload as defined by TLX

By specifying how to measure workload, TLX implicitly defines workload. Other workload measures contain other implicit definitions, but the widespread use of TLX has made its definition of workload influential. Using TLX implies defining workload as experienced, general, and accumulated.

As a self-report instrument, TLX measures experienced – or perceived – workload. This way, workload is in the eye of the beholder; it is a genuinely user-centered notion. The field of human-computer interaction has a long history of attending to such notions, including satisfaction, technology acceptance, and user experience. In other fields, such as human factors and ergonomics, the intangibility of self-reported workload has caused debate (de Winter, 2014; Dekker & Hollnagel, 2004; Parasuraman et al., 2008). It should be noted that alternative measures of workload (e.g., expert assessment and secondary task performance) change the definition of workload by rejecting the premise that workload is in the eye of the beholder. Hart and Staveland (1988) acknowledge two possible sources of noise in TLX measurements: (a) users may find it difficult to translate their experience of workload into an overt evaluation and (b) the TLX instrument may lack sensitivity to experimental manipulations or psychological processes. The first source of noise is common to all self-report measures. The present review – especially the answer to the third research question – indicates that the second source of noise is not a major concern for TLX.

With its six subscales, the composite TLX score measures general workload. The individual subscales measure distinct workload dimensions. Other instruments for measuring self-reported workload have a different profile. For example, the Subjective Workload Assessment Technique (SWAT; Reid & Nygren, 1988) consists of the three dimensions time load, mental effort load, and psychological stress load. Compared to the general workload measurement obtained with TLX, Nygren (1991) contends that SWAT has a more cognitive and psychological focus. At least, it leaves out physical demand and experienced performance. Relatedly, the Workload Profile (Tsang & Velazquez, 1996) consists of eight dimensions (perceptual/central processing, response processing, spatial processing, verbal processing, visual processing, auditory processing, manual responses, and speech responses). Compared to TLX, the Workload Profile is more detailed about demands and effort but leaves out performance and frustration.

As a measure of task load, TLX measures the workload experienced over a period of time. Typically, this period has a duration of minutes, rather than seconds or hours (Hancock & Matthews, 2019). Physiological indicators can measure workload at the millisecond (e.g., electroencephalograms) and second (e.g., heart rate) levels and thus provide information about instantaneous workload. In contrast, TLX is assumed to convey the accumulated or average workload over the period spanned by the task (Xie & Salvendy, 2000). However, this assumption simplifies matters a bit because TLX measurements, like other experience ratings, are subject to peak-end effects. That is, peaks in workload appear to bias TLX measurements, especially if the peak occurs at the end of the task (Peterson & Kozhokar, 2017; Qiao et al., 2021). Probably, workload is more memorable when it is most potent (peak) and most recent (end), thereby driving TLX ratings upward compared to the average workload over the task. The peak-end effect reiterates that TLX measures experienced workload.

5.3 Limitations

Four limitations should be remembered in interpreting the results of this review. First, only one source (Google Scholar) was searched for papers to include in the review and only one person (the author) coded these papers. Additional sources and coders would provide for validating the selection and coding of the papers. That said, the 384 included papers are a sizable data set. Second, this review is not a validation of TLX against independent workload measurements. Existing studies have investigated the association between TLX and several independent workload measurements, including secondary task performance (e.g., de Winter et al., 2016) and physiological indicators such as heart rate (e.g., Widyanti et al., 2013). It is left for future research to review these associations across a large set of studies. Third, this review assigns primacy to the subscales at the expense of the composite TLX score. Specifically, studies were excluded if they reported a test of the composite TLX score, but not of the six subscales. Future research may complement the present review with one that assigns primacy to the composite TLX score. Such a review should include studies that apply raw TLX as well as studies that apply weighted TLX. Fourth, the classification of the independent variable in the tests into user, task, system, and context is somewhat coarse-grained. For example, TLX varies across domains, such as driving, healthcare, and leisure (Hertzum, 2021). This variation may influence the association between TLX and other variables but is not investigated in this review. Future research may also consider distinguishing among different types of system to investigate whether they are associated with workload in different ways or to different extents.

6 Conclusion

TLX is a widely used measure of workload, which is an important explanatory variable in human-computer interactions. To understand the qualities of TLX and its relations to other variables, this study has reviewed the associations among the TLX subscales and between TLX and variables concerning performance, user experience, and situational characteristics. The results confirm that TLX is sensitive to situational characteristics and that it helps explain the interactions among task demands, user capabilities, and performance. In summary, the results are that:

- The six subscales measure associated, but somewhat independent, dimensions of workload.
- People compensate for more demanding conditions by putting in more effort and, as a result, sometimes avoid a drop in performance.
- Differences in workload are associated with differences in error rate, task completion time, and user experience but the strength of association is merely slight to fair.
- Differences in opposite directions between workload and either error rate, task completion time, or user experience are few but occur for all subscales.
- Differences in workload dimensions are more often associated with differences in tasks and contexts than in users and systems.
- The subscales, not just the composite TLX score, are widely used for testing cross-system differences in workload.

The results are derived from a large corpus of studies and extend our understanding of TLX and workload. It is hoped that they will assist in interpreting TLX measurements and motivate future work on how workload – overload as well as underload – influences human-computer interactions.

Disclosure statement

The author has no conflicts of interest to declare.

Funding

This study has not received external funding.

References

Abich, J., Reinerman-Jones, L., & Matthews, G. (2017). Impact of three task demand factors on simulated unmanned system intelligence, surveillance, and reconnaissance operations. *Ergonomics, 60*(6), 791-809. https://doi.org/10.1080/00140139.2016.1216171

Barnes, C.M., & Van Dyne, L. (2009). 'I'm tired': Differential effects of physical and emotional fatigue on workload management strategies. *Human Relations, 62*(1), 59-92. https://doi.org/10.1177/0018726708099518

Bessière, K., Newhagen, J.E., Robinson, J.P., & Shneiderman, B. (2006). A model for computer frustration: The role of instrumental and dispositional factors on incident, session, and post-session frustration and mood. *Computers in Human Behavior, 22*(6), 941-961. https://doi.org/10.1016/j.chb.2004.03.015

Bier, L., Emele, M., Gut, K., Kulenovic, J., Rzany, D., Peter, M., & Abendroth, B. (2019). Preventing the risks of monotony related fatigue while driving through gamification. *European Transport Research Review, 11*, article 44. https://doi.org/10.1186/s12544-019-0382-4

Blane, A., Falkmer, T., Lee, H.C., & Willstrand, T.D. (2018). Investigating cognitive ability and self-reported driving performance of post-stroke adults in a driving simulator. *Topics on Stroke Rehabilitation, 25*(1), 44-53. https://doi.org/10.1080/10749357.2017.1373929

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P.W. Jordan, B. Thomas, B.A. Weerdmeester, & I.L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189-194). London: Taylor & Francis.

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

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Dai, L., Sears, A., & Goldman, R. (2009). Shifting the focus from accuracy to recallability: A study of informal note-taking on mobile information technologies. *ACM Transactions on Computer-Human Interaction, 16*(1), article 4. https://doi.org/10.1145/1502800.1502804

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

de Winter, J.C.F., Stanton, N.A., Price, J.S., & Mistry, H. (2016). The effects of driving with different levels of unrealiable automation on self-reported workload and secondary task performance. *International Journal of Vehicle Design, 70*(4), 297-324. https://doi.org/10.1504/IJVD.2016.076736

Dekker, S., & Hollnagel, E. (2004). Human factors and folk models. *Cognition, Technology & Work, 9*(2), 79-86. https://doi.org/10.1007/s10111-003-0136-9

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

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

Gao, Q., Wu, M., & Zhu, B. (2018). The influence of culture on vigilance performance and subjective experience. In *Proceedings of the EPCE2018 Conference on Engineering Psychology and Cognitive Ergonomics* (pp. 19-31). Cham: Springer. https://doi.org/10.1007/978-3-319-91122-9_2

Gawron, V.J. (2019). *Human performance, workload, and situational awareness measures handbook, Third Edition*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/9780429019562

Goode, N., Lenné, M.G., & Salmon, P. (2012). The impact of on-road motion on BMS touch screen device operation. *Ergonomics, 55*(9), 986-996. https://doi.org/10.1080/00140139.2012.685496

Gould, K.S., Røed, B.K., Saus, E.-R., Koefoed, V.F., Bridger, R.S., & Moen, B.E. (2009). Effects of navigation method on workload and performance in simulated high-speed ship navigation. *Applied Ergonomics, 40*(1), 103-114. https://doi.org/10.1016/j.apergo.2008.01.001

Gürkök, H., Hakvoort, G., & Poel, M. (2011). Evaluating user experience in a selection based brain-computer interface game: A comparative study. In *ICEC2011: Proceedings of the International Conference on Entertainment Computing* (Vol. LNCS 6972, pp. 77-88). Berlin: Springer. https://doi.org/10.1007/978-3-642-24500-8_9

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

Hancock, P.A., & Warm, J.S. (1989). A dynamic model of stress and sustained attention. *Human Factors, 31*(5), 519-537. https://doi.org/10.1177/001872088903100503

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

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction, 19*(4), 319-349. https://doi.org/10.1207/s15327051hci1904_2

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. In *Mensch & Computer 2003* (pp. 187-196): Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-80058-9_19

Hertzum, M. (2021). Reference values and subscale patterns for the task load index (TLX): A meta-analytic review. *Ergonomics, 64*(7), 869-878. https://doi.org/10.1080/00140139.2021.1876927

Hertzum, M., & Holmegaard, K.D. (2013). Thinking aloud in the presence of interruptions and time constraints. *International Journal of Human-Computer Interaction, 29*(5), 351-364. https://doi.org/10.1080/10447318.2012.711705

Hockey, G.R.J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetic framework. *Biological Psychology, 45*(1-3), 73-93. https://doi.org/10.1016/S0301-0511(96)05223-4

Hoonakker, P., Carayon, P., Gurses, A.P., Brown, R., Khunlertkit, A., McGuire, K., & Walker, J.M. (2011). Measuring workload of ICU nurses with a questionnaire survey: The NASA task load index (TLX). *IIE Transactions on Healthcare Systems Engineering, 1*(2), 131-143. https://doi.org/10.1080/19488300.2011.609524

Jazani, R.K., Miandashti, R., Kavousi, A., & Minaei, M.S. (2016). The effect of hot and humid weather on the level of mental workload among managers and supervisors on a project of South Pars phases, Iran. *Cognition, Technology & Work, 18*(1), 11-17. https://doi.org/10.1007/s10111-015-0342-2

Katsuragawa, K., Kamal, A., & Lank, E. (2017). Effect of motion-gesture recognizer error pattern on user workload and behavior. In *IUI2017: Proceedings of the 22nd International Conference on Intelligent User Interfaces* (pp. 439-449). New York: ACM Press. https://doi.org/10.1145/3025171.3025234

Kern, F., Winter, C., Gall, D., Käthner, I., Pauli, P., & Latoschik, M.E. (2019). Immersive virtual reality and gamification within procedurally generated environments to increase motivation during gait rehabilitation. In *Proceedings of the 26th IEEE Conference on Virtual Reality and 3D User Interfaces* (pp. 500-509). Piscataway, NJ: IEEE Press. https://doi.org/10.1109/VR.2019.8797828

Kratz, S., Brodien, I., & Rohs, M. (2010). Semi-automatic zooming for mobile map navigation. In *Proceedings of the MobileHCI2010 Conference on Human Computer Interaction with Mobile Devices and Services* (pp. 63-71). New York: ACM Press. https://doi.org/10.1145/1851600.1851615

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Lazar, J., Feng, J.H., & Hochheiser, H. (2017). *Research methods in human-computer interaction* (2nd ed.). Cambridge, MA: Morgan Kaufmann.

Loup-Escande, E., Frenoy, R., Poplimont, G., Thouvenin, I., Gapenne, O., & Megalakaki, O. (2017). Contributions of mixed reality in a calligraphy learning task: Effects of supplementary visual feedback and expertise on cognitive load, user experience and gestural performance. *Computers in Human Behavior, 75*, 42-49. https://doi.org/10.1016/j.chb.2017.05.006

Majrashi, K. (2019). Post-transitioning user performance on cross-device menu interfaces. *International Journal of Human-Computer Studies, 130*, 130-151. https://doi.org/10.1016/j.ijhcs.2019.06.001

Matthews, G., & Campbell, S.E. (2009). Sustained performance under overload: Personality and individual differences in stress and coping. *Theoretical Issues in Ergonomics Science, 10*(5), 417-442. https://doi.org/10.1080/14639220903106395

McKendrick, R.D., & Cherry, E. (2018). A deeper look at the NASA TLX and where it falls short. In *Proceedings of the Human Factors and Ergonomics Society 2018 Annual Meeting* (pp. 44-48). Santa Monica, CA: HFES. https://doi.org/10.1177/1541931218621010

Melman, T., de Winter, J.C.F., & Abbink, D.A. (2017). Does haptic steering guidance instigate speeding? A driving simulator study into causes and remedies. *Accident Analysis and Prevention, 98*, 372-387. https://doi.org/10.1016/j.aap.2016.10.016

Nicol, E., Komninos, A., & Dunlop, M.D. (2016). A participatory design and formal study investigation into mobile text entry for older adults. *International Journal of Mobile Human Computer Interaction, 8*(2), article 2. https://doi.org/10.4018/IJMHCI.2016040102.oa

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

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making, 2*(2), 140-160. https://doi.org/10.1518/155534308X284417

Partala, T., & Salminen, M. (2012). User experience of photorealistic urban pedestrian navigation. In *Proceedings of the AVI2012 International Working Conference on Advanced Visual Interfaces* (pp. 204-207). New York: ACM Press. https://doi.org/10.1145/2254556.2254593

Peterson, D.A., & Kozhokar, D. (2017). Peak-end effects for subjective mental workload ratings. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1), 2052-2056. https://doi.org/10.1177/1541931213601991

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

Prilla, M., Janssen, M., & Kunzendorff, T. (2019). How to interact with augmented reality head mounted devices in care work? A study comparing handheld touch (hands-on) and gesture (hands-free) interaction. *AIS Transactions on Human-Computer Interaction, 11*(3), 157-178. https://doi.org/10.17705/1thci.00118

Proctor, R.W., Wang, D.Y., & Pick, D.F. (1998). An empirical evaluation of the SYNWORK1 multiple-task work environment. *Behavior Research Methods, Instruments, & Computers, 30*(2), 287-305. https://doi.org/10.3758/BF03200657

Qiao, H., Zhang, J., Zhang, L., Li, Y., & Loft, S. (2021). Exploring the peak-end effects in air traffic controllers' mental workload ratings. *Human Factors*. https://doi.org/10.1177/0018720821994355

Rasmussen, R., & Hertzum, M. (2013). Visualizing the application of filters: A comparison of blocking, blurring, and colour-coding whiteboard information. *International Journal of Human-Computer Studies, 71*(10), 946-957. https://doi.org/10.1016/j.ijhcs.2013.06.002

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

Sharma, V., Simpson, R., LoPresti, E., & Schmeler, M. (2012). Driving backwards using a semi-autonomous smart wheelchair system (DSS): A clinical evaluation. *Applied Bionics and Biomechanics, 9*(4), 347-365. https://doi.org/10.3233/ABB-2011-0012

Smith, L., Totterdell, P., & Folkard, S. (1995). Shiftwork effects in nuclear power workers: A field study using portable computers. *Work and Stress, 9*(2-3), 235-244. https://doi.org/10.1080/02678379508256559

Son, J., & Lee, G. (2017). Comparison of two target selection methods for two-thumb touchpad typing. *International Journal of Human-Computer Interaction, 33*(10), 799-812. https://doi.org/10.1080/10447318.2017.1286810

Thapa, R.B., Ferati, M., & Giannoumis, G.A. (2017). Using non-speech sounds to increase web image accessibility for screen-reader users. In *Proceedings of the SIGDOC2017 Conference on the Design of Communication* (paper 19). New York: ACM Press. https://doi.org/10.1145/3121113.3121231

Tran, C.C., Hoang, S., Hoang, T.D., & Nguyen, T.H.H. (2018). Assessing effectiveness of wood spray painting system model for undergraduate engineering education. *Computer Applications in Engineering Education, 27*(1), 29-37. https://doi.org/10.1002/cae.22054

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

van Schaik, P., & Ling, J. (2008). Modelling user experience with web sites: Usability, hedonic value, beauty and goodness. *Interacting with Computers, 20*(3), 419-432. https://doi.org/10.1016/j.intcom.2008.03.001

Wang, Y., Zhang, W., & Salvendy, G. (2010). Effects of a simulation-based training intervention on novice drivers' hazard handling performance. *Traffic Injury Prevention, 11*(1), 16-24. https://doi.org/10.1080/15389580903390631

Widyanti, A., de Waard, D., Johnson, A., & Mulder, B. (2013). National culture moderates the influence of mental effort on subjective and cardiovascular measures. *Ergonomics, 56*(2), 182-194. https://doi.org/10.1080/00140139.2012.748219

Wilson, K.M., Helton, W.S., de Joux, N.R., Head, J.R., & Weakley, J.J.S. (2017). Real-time quantitative performance feedback during strength exercise improves motivation, competitiveness, mood, and performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*, 1546-1550. https://doi.org/10.1177/1541931213601750

Xie, B., & Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. *Work and Stress, 14*(1), 74-99. https://doi.org/10.1080/026783700417249

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

Yu, Y., & Liu, Z. (2010). Improving the performance and usability for visual menu interface on mobile computers. In *Proceedings of the AVI2010 International Conference on Advanced Visual Interfaces* (pp. 369-372). New York: ACM Press. https://doi.org/10.1145/1842993.1843065

Appendix A

Cross-tabulation of the TLX subscales against error rate, task completion time, and user experience. The cross-tabulations only include the tests for which data were available for the subscale as well as for error rate, task completion time, or user experience. Each cross-tabulation gives:

NS: the number of tests that did not find a significant difference between the two compared conditions

C1: the number of tests that found a significant difference in favor of Condition 1 (in terms of lower workload, lower error rate, lower task completion time, or higher user experience).

C2: the number of tests that found a significant difference in favor of Condition 2.

	Error rate			Task completion time			User experience		
	NS	C1	C2	NS	C1	C2	NS	C1	C2
Mental demand									
NS	174	50	42	170	81	61	148	52	41
C1	31	14	1	22	39	5	22	32	2
C2	40	4	43	29	4	47	24	9	35
Physical demand									
NS	183	48	49	165	88	66	155	69	42
C1	20	13	1	18	28	5	13	16	2
C2	37	7	36	38	8	42	26	8	34
Temporal demand									
NS	192	55	49	175	92	70	156	65	49
C1	15	8	3	15	21	3	15	18	0
C2	38	5	34	30	10	40	23	10	29
Effort									
NS	174	40	45	173	76	55	147	54	36
C1	28	23	1	19	43	7	21	33	4
C2	43	5	40	29	5	51	26	6	38
Performance									
NS	213	45	56	184	91	76	167	61	41
C1	12	21	3	19	26	5	12	23	8
C2	20	2	27	17	6	32	15	9	29
Frustration									
NS	183	50	48	175	90	58	165	60	43
C1	21	15	0	14	29	5	11	26	3
C2	41	3	38	31	4	50	18	7	32

TLX									
NS	85	19	20	86	34	23	87	16	10
C1	13	18	1	16	25	5	9	20	0
C2	27	3	30	15	5	26	16	10	18

Table 1. The six TLX subscales

Subscale	Definition [a]
Mental demand	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	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	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	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Performance	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	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

[a] Cited from Hart and Staveland (1988).

Table 2. The classification of the independent variable in the tests

Category	Description
User	Tests comparing workload across groups of users. For example, Nicol et al. (2016) compare the workload experienced by older and younger adults during mobile text entry, Loup-Escande et al. (2017) compare the workload experienced by novices and experts in a calligraphy task on a tablet, and Gao et al. (2018) compare the workload experienced by American and Chinese students on a vigilance task.
Task	Tests comparing workload across tasks. For example, Proctor et al. (1998) compare the workload imposed by arithmetic tasks and visual-monitoring tasks, Hoonakker et al. (2011) compare the workload imposed on nurses by an 8-hour or 12-hour shift, and Abich et al. (2017) compare the workload imposed by reconnaissance tasks with a low or high event rate.
System	Tests comparing workload across systems. For example, Gould et al. (2009) compare the workload imposed by electronic and paper charts during ship navigation, Yu and Liu (2010) compare the workload imposed by visual and sonically enhanced menus on mobile computers, and Gürkök et al. (2011) compare the workload imposed by brain-computer interaction and automatic speech recognition for object selection in a computer game.
Context	Tests comparing workload across contexts. For example, Goode et al. (2012) compare the workload experienced by passengers during touchscreen input in vehicles driven on sealed or unsealed roads, Hertzum and Holmegaard (2013) compare the workload imposed by a computer game when played in silence or while thinking aloud, and Jazani et al. (2016) compare the workload experienced by supervisors during summer (hot and humid) and winter (cooler).

Table 3. Tests that found a significant difference in workload versus those that did not

	Significant difference		No significant difference		Total
	N	%	*N*	%	*N*
Mental demand	311	35	590	65	901
Physical demand	276	31	623	69	899
Temporal demand	246	27	655	73	901
Effort	321	36	582	64	903
Performance	222	25	679	75	901
Frustration	280	31	621	69	901
TLX	201	44	256	56	457

Table 4. Associations among TLX and its subscales in terms of percent agreement (below diagonal) and Kappa (above diagonal), *N* is the number of tests over which the association was analyzed

	MD	PD	TD	EF	PE	FR	TLX
Mental demand (MD)		.35	.45	.57	.32	.45	.60
		N = 896	*N* = 899	*N* = 901	*N* = 899	*N* = 899	*N* = 457
Physical demand (PD)	68%		.35	.42	.24	.36	.34
	N = 896		*N* = 896	*N* = 898	*N* = 896	*N* = 896	*N* = 457
Temporal demand (TD)	74%	71%		.46	.31	.47	.47
	N = 899	*N* = 896		*N* = 901	*N* = 901	*N* = 901	*N* = 455
Effort (EF)	78%	71%	74%		.33	.49	.61
	N = 901	*N* = 898	*N* = 901		*N* = 901	*N* = 901	*N* = 457
Performance (PE)	68%	67%	71%	68%		.38	.41
	N = 899	*N* = 896	*N* = 901	*N* = 901		*N* = 901	*N* = 455
Frustration (FR)	73%	70%	76%	75%	73%		.49
	N = 899	*N* = 896	*N* = 901	*N* = 901	*N* = 901		*N* = 455
TLX	78%	65%	71%	78%	69%	72%	
	N = 457	*N* = 457	*N* = 455	*N* = 457	*N* = 455	*N* = 455	

Table 5. Association between TLX and error rate

	Agreement across all tests			Agreement across significant tests		
	%	Kappa	*N*	%	Kappa	*N*
Mental demand	58	.19	399	92	.79	62
Physical demand	59	.19	394	86	.67	57
Temporal demand	59	.16	399	84	.56	50
Effort	59	.23	399	91	.82	69
Performance	65	.27	399	91	.81	53
Frustration	59	.19	399	95	.87	56
TLX	62	.32	216	92	.84	52

Note: *%* – percent agreement in the outcome of the workload and error-rate tests, *Kappa* – Cohen's Kappa of the agreement in the outcome of the workload and error-rate tests, and *N* – the number of tests over which the association was analyzed

Table 6. Association between TLX and task completion time

	Agreement across all tests			Agreement across significant tests		
	%	Kappa	*N*	%	Kappa	*N*
Mental demand	56	.25	458	91	.81	95
Physical demand	51	.17	458	84	.68	83
Temporal demand	52	.16	456	82	.63	74
Effort	58	.30	458	89	.77	106
Performance	53	.18	456	84	.68	69
Frustration	56	.24	456	90	.78	88
TLX	58	.30	235	84	.67	61

Note: *%* – percent agreement in the outcome of the workload and completion-time tests, *Kappa* – Cohen's Kappa of the agreement in the outcome of the workload and completion-time tests, and *N* – the number of tests over which the association was analyzed

Table 7. Association between TLX and user experience

	Agreement across all tests			Agreement across significant tests		
	%	Kappa	*N*	%	Kappa	*N*
Mental demand	59	.28	365	86	.72	78
Physical demand	56	.21	365	83	.64	60
Temporal demand	56	.19	365	82	.65	57
Effort	60	.30	365	88	.75	81
Performance	60	.27	365	75	.50	69
Frustration	61	.29	365	85	.71	68
TLX	67	.41	186	79	.60	48

Note: *%* – percent agreement in the outcome of the workload and user-experience tests, *Kappa* – Cohen's Kappa of the agreement in the outcome of the workload and user-experience tests, and *N* – the number of tests over which the association was analyzed

Table 8. Breakdown of the tests that found a significant difference in workload onto those comparing user groups, tasks, systems, and contexts

	Significant difference		Total
	N	%	
User			
Mental demand	17	20	86
Physical demand	27	31	86
Temporal demand	25	29	85
Effort	23	27	86
Performance	21	25	85
Frustration	22	26	85
TLX	16	30	53
System			
Mental demand	149	31	487
Physical demand	130	27	482
Temporal demand	96	20	486
Effort	166	34	487
Performance	101	21	486
Frustration	137	28	486
TLX	95	41	232
Task			
Mental demand	77	49	158
Physical demand	53	33	160
Temporal demand	59	37	160
Effort	64	40	160
Performance	45	28	160
Frustration	62	39	160
TLX	42	48	88
Context			
Mental demand	64	41	155
Physical demand	61	39	155
Temporal demand	61	39	155
Effort	67	43	155
Performance	49	32	155
Frustration	55	35	155
TLX	43	55	78

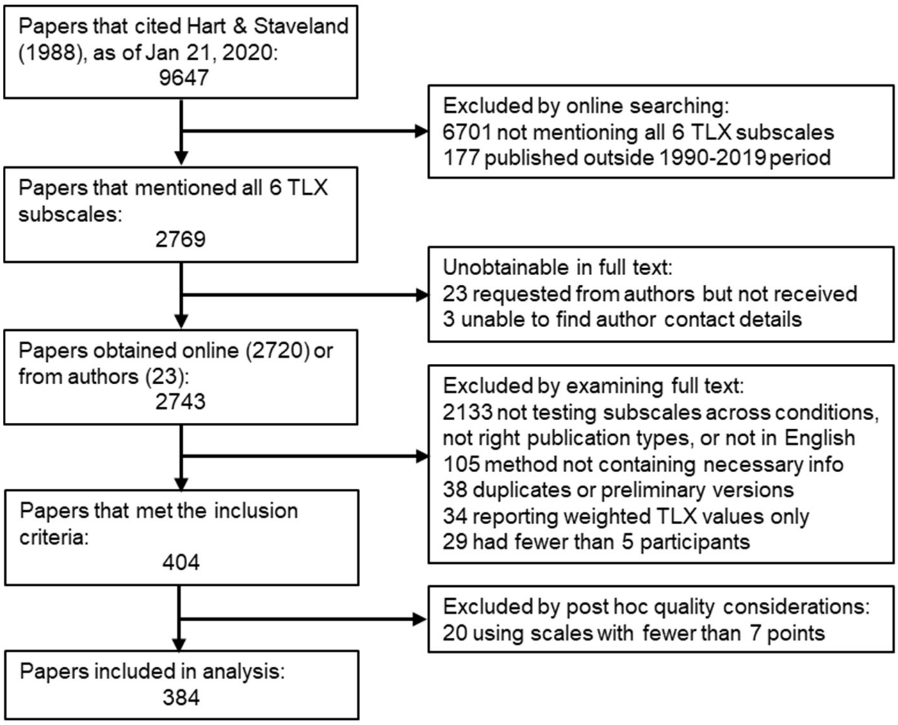
Papers that cited Hart & Staveland (1988), as of Jan 21, 2020: 9647
Excluded by online searching: 6701 not mentioning all 6 TLX subscales 177 published outside 1990-2019 period
Papers that mentioned all 6 TLX subscales: 2769
Unobtainable in full text: 23 requested from authors but not received 3 unable to find author contact details
Papers obtained online (2720) or from authors (23): 2743
Excluded by examining full text: 2133 not testing subscales across conditions, not right publication types, or not in English 105 method not containing necessary info 38 duplicates or preliminary versions 34 reporting weighted TLX values only 29 had fewer than 5 participants
Papers that met the inclusion criteria: 404
Excluded by post hoc quality considerations: 20 using scales with fewer than 7 points
Papers included in analysis: 384

Figure 1. Paper-selection process