Context Sensitive Health Informatics: The Role of Informatics in Global Pandemics
R. Marcilly et al. (Eds.)

doi:10.3233/SHTI210637

How Demanding Is Healthcare Work? A Meta-Analytic Review of TLX Scores

Morten HERTZUM [a,1]
[a] *University of Copenhagen, Denmark*

Abstract. This study establishes how demanding healthcare work is experienced to be and whether nurses and physicians experience different levels of workload. A meta-analytic review was conducted of 87 studies that reported Task Load Index (TLX) scores for healthcare work. Of these studies, 37 were conducted in real-life settings and 50 in lab settings without real patients. In real-life settings, clinicians experienced a workload with a mean TLX of 49 (on a 0-100 scale). Divided onto staff groups, the mean TLX for nurses was 63, which was significantly higher than the mean of 40 for physicians. Among the six TLX subscales, the main contributors to workload were mental demand, temporal demand, and effort. They were higher than physical demand and frustration. The clinicians experienced their performance – the last subscale – as closer to poor than good in 38% of the studies conducted in real-life settings. The difference between nurses and physicians was consistent across all subscales, except mental demand. Finally, it is methodologically important that TLX scores appeared not to transfer directly from lab to real-life settings. To reduce the risk of errors and burnout, new healthcare procedures and technologies should be evaluated for their impact on workload.

Keywords. Workload, psychological stress, physicians, nurses, NASA-TLX

1. Introduction

Excessive workload increases the risk of errors and burnout, thereby harming patients as well as healthcare staff [1, 2]. Accordingly, it is important to know the workload imposed by healthcare work and how it is experienced by physicians, nurses, and other staff. Without such knowledge, risks may go unnoticed. With it, the high-workload areas can become the target of efforts to reduce workload. For example, electronic whiteboards are often introduced to reduce workload by improving communication and overview [3, 4]. This study establishes healthcare clinicians' workload by reviewing published Task Load Index (TLX, aka NASA-TLX) scores.

TLX [5] measures self-reported workload. Self-reported workload is important because clinicians who experience their workload as excessive will behave as though they are overloaded, irrespective of the objective task demands. Among the measures of workload, TLX is so widely used that de Winter [6, p. 293] has stated that "workload has become synonymous with the TLX." Workload emerges from the interaction between the demands imposed by work tasks and the skills, behaviors, and perceptions of the staff performing the work [5]. A TLX score is the mean of six subscales: mental demand, physical demand, temporal demand, effort, performance, and frustration. Each subscale

[1] Corresponding Author, Morten Hertzum, Department of Communication, University of Copenhagen, Karen Blixens Plads 8, 2300 Copenhagen, Denmark; E-mail: hertzum@hum.ku.dk.

is measured with a single item that has the endpoints 'Low' (0) and 'High' (100), except that performance has the endpoints 'Good' (0) and 'Poor' (100).

This study reviews published TLX scores to establish the workload experienced by clinicians. To obtain information about the different dimensions of workload, the review is restricted to studies that include values for the six TLX subscales. The study seeks to answer two research questions:

- How demanding is healthcare work experienced to be?
- Do nurses and physicians experience different levels of workload?

Furthermore, the study distinguishes between workload measurements obtained in real-life settings and in lab settings (i.e., in experiments that do not involve real patients). This distinction helps avoid undue transfer of TLX scores between the two settings.

2. Method

The authoritative reference for TLX is Hart and Staveland [5]. Thus, the three primary inclusion criteria for this study were that papers cited Hart and Staveland [5], were about healthcare, and reported empirical values for all six TLX subscales. The values had to be raw TLX ratings; papers that reported weighted ratings for the subscales were excluded because weighting was infrequent and has been depreciated [7]. The papers also had to have at least five participants, be published in journals, edited books, or conference proceedings in the period 1990-2019, and be in English. When a paper existed in multiple versions, only the most extensive version was included. Initially, Google Scholar was searched for the papers that cited Hart and Staveland [5]. Of these 9647 papers, 86 met the inclusion criteria. One paper reported from 2 studies, for a total of 87 studies.

The data analysis had four steps. First, the papers were coded. This involved extracting the number of study participants, the numerical endpoints of the rating scales, the participant group, the setting (lab or real life), and the values of the TLX subscales. Subscale values were extracted for each condition for which such values were reported. Second, the subscale values were rescaled to the 0-100 range, if another range was used in the study. Third, TLX was calculated as the mean of the subscales. Fourth, a value for each study was obtained by taking the mean across the study conditions, thereby making the analysis independent of the number of conditions in the studies. Each of the 87 studies contributed to the analysis with a score for each subscale and a TLX score.

3. Results

Of the 87 studies, 37 reported from studies conducted in real-life settings. These 37 studies involved 5739 participants. The remaining 50 studies were conducted in lab settings and involved 1454 participants. Figure 1 shows the distribution of TLX. The median TLX score was 44 (lab) and 47 (real life), whereas the mean TLX score was 42 (lab) and 49 (real life). Table 1 shows the distribution of the six TLX subscales for the studies conducted in *real-life settings*. One in five studies found that the workload was 70 or more for mental demand, temporal demand, effort, performance, and overall TLX. In 38% of the studies, performance was closer to poor than good. The subscales differed significantly, $F(5, 32) = 11.76$, $p < .001$. Bonferroni-adjusted pairwise comparisons

showed that mental demand, temporal demand, and effort were higher than physical demand and frustration and that performance was lower (i.e., better) than mental demand. The subscales also differed significantly for the studies conducted in *lab settings*, $F(5, 45) = 18.00$, $p < .001$. However, the pairwise comparisons showed a different pattern for physical demand (it was lower than all other subscales), temporal demand (it was lower than mental demand and effort), and performance (it did not differ from mental demand).

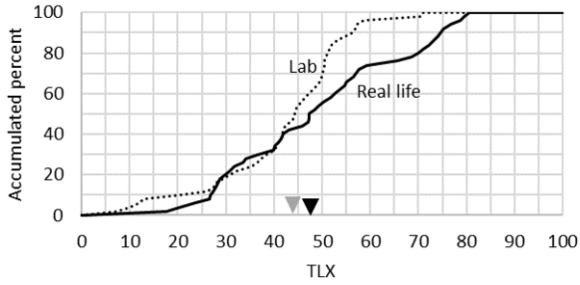

Figure 1. Accumulated distribution of TLX, $N = 50$ (lab) + 37 (real life) studies. The triangles indicate the median for lab (gray) and real-life (black) studies.

Table 1. Distribution of TLX and its subscales in real-life settings, $N = 37$ studies

Percentile	MD	PD	TD	EF	PE	FR	TLX
10th percentile	32	18	21	30	11	20	27
20th percentile	36	24	31	39	23	24	30
30th percentile	45	30	41	45	24	28	36
40th percentile	55	33	49	48	28	31	42
50th percentile (median)	62	40	56	54	39	39	47
60th percentile	63	45	63	60	45	46	53
70th percentile	70	50	67	62	70	52	57
80th percentile	77	61	70	74	77	58	70
90th percentile	81	71	76	82	80	68	74
100th percentile	88	84	87	86	88	80	80

Note: EF – effort, FR – frustration, MD – mental demand, PD – physical demand, PE – performance, TD – temporal demand, TLX – task load index

Table 2 shows the TLX data divided onto nurses, physicians, and other study participants. The 'other' group mainly consisted of lab studies with student participants. There was a significant effect of participant group on TLX for the studies conducted in *real-life setting*s, $F(2, 34) = 10.74$, $p < .001$. Bonferroni-adjusted pairwise comparisons showed that nurses (63) experienced significantly higher workload than physicians (40) and other participants (40). With nurses experiencing 58% higher TLX scores than physicians, the difference in workload was not just statistically significant but also large. There were also significant effects of participant group on mental demand, physical demand, temporal demand, effort, performance, and frustration, $F(2, 34) = 3.43$, 6.00, 7.10, 6.66, 10.54, and 11.72, respectively (all ps $< .05$). Bonferroni-adjusted pairwise comparisons showed that all subscales, save mental demand, were higher for nurses than physicians and that all subscales, save mental and physical demand, were higher for nurses than other participants. The largest subscale difference concerned performance. Physicians (30) experienced that they performed twice as well as nurses (65) did. As a result, performance drove TLX downward for the physicians, but upward for the nurses. For the *lab studies*, there was no effect of participant group on workload, $F(2, 47) = 0.24$,

0.49, 2.36, 0.02, 1.16, 2.62, and 0.24 (all *p*s > .08) for mental demand, physical demand, temporal demand, effort, performance, frustration, and TLX, respectively.

Table 2. TLX and its subscales (mean ± standard deviation) for participant groups

Group	Studies	MD	PD	TD	EF	PE	FR	TLX
Lab								
- Nurses	2	53±3	22±13	62±7	50±1	47±17	55±14	48±9
- Physicians	22	49±18	32±14	42±13	39±12	39±12	40±11	42±12
- Other	26	46±18	34±17	39±16	48±18	46±21	35±14	41±15
- Total	50	47±18	32±16	41±15	48±16	43±17	38±13	42±13
Real life								
- Nurses	14	67±18	54±20	66±14	66±16	65±21	56±15	63±14
- Physicians	16	53±15	33±12	46±16	48±11	30±17	32±11	40±9
- Other	7	49±21	34±25	39±27	47±21	38±32	33±21	40±22
- Total	37	58±18	41±20	52±20	55±17	45±27	41±19	49±18

Note: EF – effort, FR – frustration, MD – mental demand, PD – physical demand, PE – performance, TD – temporal demand, TLX – task load index

4. Discussion

Healthcare has a mean TLX of 49 in real-life settings. For comparison, the mean TLX reported from real-life settings across a range of domains is also 49 [8]. However, the workload in healthcare is not uniformly high. In 20% of the reviewed real-life studies, TLX is 70 or more. For example, Sönmez et al. [9] report a mean TLX of 80 for nurses working in a variety of hospital units. In addition, TLX is 58% higher for nurses than physicians. Physicians, for example, experienced a TLX of 48-54 (depending on their level of experience) after the introduction of an electronic whiteboard [3]. Nursing is a high-workload job, so much so that nurses' mean experience of their performance is closer to poor than good. The higher workload experienced by nurses than physicians is consistent across all subscales, save mental demand.

Staff-group differences aside, workload in real-life healthcare settings is mostly about mental demand, temporal demand, and effort. By contrast, physical demand and frustration contribute less to workload. These subscale patterns show that the demands imposed by healthcare work are mental and temporal to a larger extent than physical. Two other inferences should also be noted. First, clinicians try to compensate for high demands by expending extra effort in an attempt to maintain their level of performance. With effort as a main contributor to their workload, clinicians are straining themselves, thereby suggesting a small margin between their workload and overload. Second, even with the extra effort, 38% of the real-life studies find that clinicians experience their performance as closer to poor than good. This could indicate that many clinicians have adopted suboptimal behaviors, such as shortcuts, to cope with their workload [10]. The pattern that frustration is a modest contributor to workload suggests that the high demands are so common that clinicians have come to perceive these demands as integral to the normal state of affairs. In general, frustration ensues when events unexpectedly thwart goal attainment [11]. That is, high demands cause less frustration if they are normal to the extent of being expected. In healthcare, it appears that high workload is sufficiently common to make frustration a modest contributor to workload.

This meta-analytic review has several implications for research and practice. First, the workload in healthcare is sufficiently high to be a risk factor that increases the likelihood of errors and burnout, more so for nurses than physicians. Consistent with

previous studies [1, 2], the TLX subscales indicate that clinicians are straining themselves and may be adopting suboptimal behaviors to cope with their workload. Second, there is a need for procedures and technologies to help reduce workload in safe ways. These procedures and technologies may simplify tasks [4] or provide early warnings of possible overload [12]. Relatedly, new procedures and technologies should be evaluated for their impact on workload. If introduced without evaluation, it may go unnoticed that a staff group gets overloaded. The associated risks may dwarf the benefits of the procedure or technology. Third, the TLX scores reported in this study (Tables 1 and 2) can serve as reference values against which to evaluate local TLX measurements. Such comparison against an independent corpus may be preferable to the effort of establishing a local point of reference. Fourth, TLX scores obtained in lab and real-life settings appear to differ. Possible explanations for this difference include that some tasks may be studied more in one or the other setting. In addition, real-life settings involve more multitasking and genuine consequences [8]. The workload difference between real-life and lab settings precludes direct cross-setting comparisons and complicates the pre-implementation evaluation of how procedures and technologies affect workload.

The review results should be interpreted with certain limitations in mind. The reviewed TLX scores concern the workload perceived during selected work tasks. These tasks set the context for the measurements but have not been specified in the review, except by distinguishing between nurses' and physicians' tasks. In addition, only one source (Google Scholar) was searched for papers to include in the review and only one person (the author) coded these papers.

References

[1] J.S. Weissman, J.M. Rothschild, E. Bendavid, et al., Hospital workload and adverse events, *Medical Care* **45** (2007), 448-455.

[2] C.P. West, L.N. Dyrbye, and T.D. Shanafelt, Physician burnout: Contributors, consequences and solutions, *Journal of Internal Medicine* **283** (2018), 516-529.

[3] D.J. France, S. Levin, R. Hemphill, K. Chen, D. Rickard, R. Makowski, I. Jones, and D. Aronsky, Emergency physicians' behaviors and workload in the presence of an electronic whiteboard, *International Journal of Medical Informatics* **74** (2005), 827-837.

[4] M. Hertzum and J. Simonsen, Work-practice changes associated with an electronic emergency department whiteboard, *Health Informatics Journal* **19** (2013), 46-60.

[5] S.G. Hart and L.E. Staveland, Development of NASA-TLX (task load index): Results of empirical and theoretical research, in: *Human Mental Workload,* P.A. Hancock and N. Meshkati, eds., North-Holland, Amsterdam, 1988, pp. 139-183.

[6] J.C.F. de Winter, Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective, *Cognition, Technology & Work* **16** (2014), 289-297.

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

[8] M. Hertzum, Reference values and subscale patterns for the task load index (TLX): A meta-analytic review, *Ergonomics* **64** (2021), 869-878.

[9] B. Sönmez, Z. Oguz, L. Kutlu, and A. Yildirim, Determination of nurses' mental workloads using subjective methods, *Journal of Clinical Nursing* **26** (2016), 514-523.

[10] J.R.B. Halbesleben, D.S. Wakefield, and B.J. Wakefield, Work-arounds in health care settings: Literature review and research agenda, *Health Care Management Review* **33** (2008), 2-12.

[11] R. Yu, The neural basis of frustration state, in: *Neuroimaging Personality, Social Cognition, and Character,* J.R. Absher and J. Cloutier, eds., Academic Press, Amsterdam, 2016, pp. 223-243.

[12] K.L. Forsyth, H.J. Hawthorne, W.D. Cammon, A.R. Linden, and R.C. Blocker, Perceived workload and an automated workload alert system: A comparison in the emergency department, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **62** (2018), 573-577.