Original Article

Comparison of In-Flight and Post-Flight Use of NASA-TLX

Heikki Mansikka[1,2], Kai Virtanen[1,2], and Don Harris[3]

[1]Systems Analysis Laboratory, Aalto University, Helsinki, Finland
[2]Department of Military Technology, National Defence University, Helsinki, Finland
[3]Faculty of Engineering, Environment and Computing, Coventry University, UK

Abstract: The objective of this study was to evaluate the feasibility of airborne assignment of NASA-TLX scores. A total of 21 participants flew two simulated flying missions. In both missions, the participants' task was to fly an instrument landing system (ILS) approach, followed by a stabilized climbing turn maneuver. In one mission, the participants evaluated their mental workload on the ILS approach and assigned the NASA-TLX scores during the climbing turn maneuver. In the other mission, the same tasks were conducted after the simulated flying task. The participants' NASA-TLX scores and flying performance between missions were compared. There was no significant difference in the participants' flying performance between missions. There were no significant differences in either NASA-TLX scores or overall indices between two missions.

Keywords: flying performance, NASA-TLX, simulated flight, time delay

A pilot's balanced mental workload (MWL) during flight is a critical enabler for operational effectiveness and flight safety (Aasman et al., 1987; Mansikka et al., 2019; Paas & Van Merriënboer, 1993). The National Aeronautics and Space Administration – Task Load Index (NASA-TLX; Hart, 2006; Hart & Staveland, 1988) is a commonly used method to asses pilots' MWL (see, e.g., Casali & Wierwille, 1984; Jorna, 1993; Mansikka et al., 2016; Sohn & Jo, 2003; Wilson, 1993). It is frequently used as a simple way to obtain a post-task estimate of the subjective MWL. NASA-TLX is a multidimensional method, where MWL is assessed across six dimensions: mental demand (MD), physical demand (PD) and temporal demand (TD), which reflect the sources of workload, and performance (OP), effort (EF), and frustration level (FR), which represent the interaction of the pilot's performance with the task. MWL assessments conducted across the dimensions have been shown to have reasonably high levels of test/retest reliability (Corwin, 1989). Corwin reports slightly higher levels of test/retest reliability for the NASA-TLX dimensions that reflect sources of workload (correlation typically at r = .40–.80) than for those that represent the pilot's interaction with the task (correlation typically in the region of r = .35–.75).

NASA-TLX, like all subjective MWL measures, is low cost, nonintrusive, and easy to administer. A literature search on Google Scholar suggests that since 2000, NASA-TLX has been used in almost 10,000 published works in the aviation and aerospace domain. The measure has been used for various purposes, for example, to support the flight deck design process (e.g., Banks et al., 2018; Li & Chen, 2020), to examine the workload imposed by different in-flight failures (Etherington et al., 2016) or procedures (Efthymiou et al., 2019), and to validate or corroborate other measures of workload. Such measures include heart rate (e.g., Alaimo et al., 2021; Mansikka et al., 2018, 2019), bilateral prefrontal cortex blood oxygenation using functional near-infrared spectroscopy (e.g., Bishop et al., 2021; Matthews et al., 2015), pupillometry (Silva et al., 2021), and respiratory sinus arrhythmia (Muth et al., 2012).

However, despite its popularity, NASA-TLX has several limitations. Annett (2002a) outlined general concerns relating to the validity and reliability of subjective workload measures (for an in-depth discussion about the topic, see Annett, 2002a, 2002b; Young & Stanton, 2002). As Annett (2002b) has pointed out, any time delay between a task execution and the person's retrospective evaluation of his/her MWL when doing that task is a source of systematic time error. In a study of airline pilots flying simulated missions in a Boeing 727, Corwin (1992) observed that post-flight scores of high workload events using another subjective MWL measure, the Subjective Workload Analysis Technique (SWAT; Reid & Nygren, 1988), were significantly higher than in-flight scores. Moroney et al. (1995) conducted an experimental investigation of the effects of delay when making workload measurements using SWAT

https://doi.org/10.1027/2192-0923/a000272

and NASA-TLX, both in the laboratory and in simulated flight. When performing a compensatory tracking task, it was observed that the greater the delay before providing the workload scores with SWAT, the more the scores between high- and low-workload experimental scenarios converged (the range became more restricted). The interference effect of a working memory task that required participants to memorize the status of four different types of information (see Eggemeier et al., 1983), between workload evaluation task and workload scores was negligible (up to a 30-min delay). Moroney et al. (1995) reported similar results when making workload scoring using NASA-TLX, although the maximum meaningful delay before making the workload assessments in this case was only 15 min.

By contrast, Tsang and Vidulich (1994) evaluated the impact of time delay when using three different subjective workload techniques: immediate workload assessment (using a simple, univariate 0–100 scale); delayed workload assessment (using the same scale); and Subjective Workload Dominance (SWORD) ratings (Vidulich, 1989). They proposed that by making ratings immediately after the task of interest, the potential for workload misinterpretations arising from memory decay would be minimized. This would suggest that by minimizing the time delay between the task execution and the assignment of the NASA-TLX workload scores, the validity of the workload measures is enhanced.

When taken to an extreme, the NASA-TLX scores could be given airborne, right after the task of interest but while still flying. However, an evaluation of workload and assignment of scores during flight can also potentially increase the pilot's task demand. This increase has potential to lead to an elevated MWL and degraded performance. If this increases demands on working memory above a critical point, it may also lead to elevated pilot's MWL and decrease pilot's accuracy in controlling an aircraft. Harris (2011) commented that it is unlikely that SWAT or NASA-TLX could be used to make concurrent workload estimates during simulated or real flight as a result of the number of subscale scores required. Both workload approaches are likely to intrude on task performance. In addition, the increased task demand during workload evaluation and the assignment of NASA-TLX scores could affect the pilots' opinions about their MWL in the task that they are supposed be assessing. Moreover, the impact an additional task of in-flight administration of NASA-TLX has on flying performance has not been explored. In this study, a simulated flying mission was used to investigate these issues.

To evaluate whether the airborne evaluation of workload and the assignment of NASA-TLX scores are feasible, it is necessary to investigate how much the additional task of assigning the scores impacts the pilots' MWL and control accuracy, and also whether it affects the NASA-TLX scores themselves. Such an analysis, however, creates an interesting research question. Traditionally, Type I error is the error to avoid. In this instance, what is actually of interest is whether completing NASA-TLX in simulated flight makes no difference to workload scores or control accuracy – an issue requiring consideration of Type II error (see, e.g., Harris, 1991). A nonsignificant result when comparing NASA-TLX scores between these conditions does not necessarily indicate that they are the same: It just suggests that they are not significantly different. However, Type II error is concerned with the probability of failing to observe a significant difference where one actually exists. In this study where the focus is on establishing whether there is no difference between in-flight and post-flight scores of workload, Type II error is equally important as Type I error. In this study, the NASA-TLX scores obtained during and after the flying mission were compared. In addition, the impact the in-flight assignment of NASA-TLX scores has on pilots' control accuracy was evaluated.

Method

Participants

A total of 21 cadets from the Finnish Air Force Academy volunteered to participate in the study. The mean age of the participants was 21.33 years (*SD* = 0.94). All participants were male. The participants were at the same phase in their flying curriculum and their mean experience on a Valmet L-70 Vinka aircraft was 70.66 hr (*SD* = 19.13), including their experience on an aircraft's flight training device (FTD). Written informed consent was obtained from each participant before the trial. At the time of the trial, all participants were medically fit to fly and were qualified to fly the flying mission used in the trial. The flying mission was part of the participants' normal flying duties. The missions were flown during their normal office hours.

In Finland, ethical review of nonmedical research involving human participants is based on a set of guidelines drawn up by the Finnish National Board on Research Integrity. According to its guidelines, the research configuration of this paper was such that it did not require an ethical review statement from a human sciences ethics committee.

Apparatus

The Valmet L-70 Vinka is a two-seat aircraft used for elementary flying training by the Finnish Air Force. A Vinka FTD, which simulates the Valmet L-70 Vinka, was used for the flying task. The FTD has a 180-degree high-fidelity visual system and a fully functional cockpit, which

Figure 1. Vinka flight training device. The instructor's operation station is behind the pilot's seat.

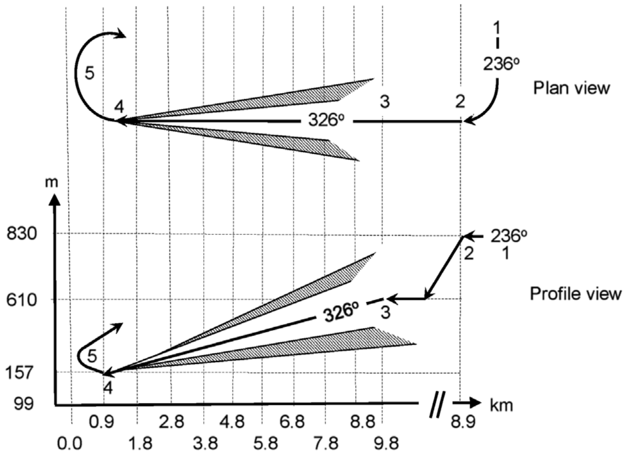

Figure 2. Instrument landing system approach procedure used in missions. A plan view of the approach profile is shown on the top and a profile view on the bottom. Headings are magnetic. Figure is not illustrated to scale.

accurately replicates that of the aircraft. The flight instructor can manipulate the environmental conditions and the FTD's systems via the instructor's operation station (see Figure 1).

Flying Missions

Each simulation trial consisted of two flying missions: an in-flight NASA-TLX mission and a post-flight NASA-TLX mission. The missions differed only in the timing when workload was evaluated and NASA-TLX scores were assigned. For each flying mission, wind was initially set at 15 kn from heading 020 degrees. Solid cloud was set from 99 to 2,000 m above mean sea level (MSL). The visibility inside the cloud was set at 0 m. The participants' task was to fly two instrument landing system (ILS) approaches, one in each mission, to Kuopio (EFKU) aerodrome according to an ILS Z runway 33 instrument approach chart published by the Air Navigation Services Finland (the instrument approach chart is available from: https://www.ais.fi/ais/aip/ad/efku/EF_AD_2_EFKU_33_ILSZ.pdf). ILS is a precision approach procedure, which provides a pilot with both horizontal and vertical control cues throughout the approach profile. There was a 3/min rest period between the two missions. The task demand of the ILS approach in both missions was identical.

The FTD was initialized at 830 m MSL (Figure 2, Point 1) flying straight and level at 180 km/h, heading 250 degrees. The participants' task was to fly over an initial approach fix (see Figure 2, Point 2), descend to 610 m MSL, and intercept the ILS localizer by the time they reached the final approach fix (see Figure 2, Point 3). At the final approach fix, the participants intercepted the ILS glide slope. Once the localizer and the glide slope had been intercepted, the participants' task was to fly the ILS approach by minimizing the horizontal and vertical control errors (with respect to the control cues) until they reached the 157-m MSL, which was the decision height (DH) for the ILS Z 33 approach (Figure 2, Point 4). When the participants passed 550 m at approximately 8.5 km from DH, the direction of the wind was changed from 020 degrees to 310 degrees. As the participants descended through 330 m at approximately 3.5 km from DH, the direction of the wind was changed back to 020 degrees. With a standard approach speed of 180 km/h and a descent rate of 2.51 m/s, the ILS approach from glideslope interception to DH took 3 min. After reaching DH, the pilots were to commence a go-around, retract flaps, and fly a stabilized right climbing turn for 3 min (Figure 2, Point 5). During this climbing turn maneuver, the participants were to maintain an airspeed of 140 km/h at full power and a bank angle of 15 degrees.

Measures

NASA-TLX

When NASA-TLX is administered, participants provide two types of information about each dimension: weights and scores. The weights represent the subjective importance of each dimension as a source of MWL during the task of interest, whereas the scores express the subjectively experienced magnitude of MWL with respect to each dimension. A weighted score for each dimension is calculated by multiplying the dimension's score by its normalized weight. Here, normalization means that the sum of weights is 1. An overall MWL index is the sum of the weighted dimension scores.

Before the trials, the participants weighted the NASA-TXL dimensions with respect to their importance as a source of MWL on a typical EFKU ILS Z33 approach task. As Virtanen et al. (2021) discussed and demonstrated, the original pairwise comparison procedure used in NASA-TLX for obtaining the dimensions' weights has a few challenges. To overcome these, the Swing weighting method (Von Winterfeldt & Edwards, 1986) was used. In this method, the most important dimension (or dimensions) is (are) assigned 100 points. The less important dimensions are assigned less than 100 points with respect to their relative importance. Unlike the original NASA-TLX weighting procedure, Swing allows for dimensions to be expressed as equally important and it does not artificially limit weight values.

In the in-flight NASA-TLX mission, the participants evaluated workload and scored the NASA-TLX dimensions regarding the ILS approach flown immediately prior to making the scoring (Figure 2, from Point 1 to Point 4) while flying the climbing turn maneuver (Figure 2, Point 4 onwards). In the post-flight NASA-TLX mission, the participants evaluated workload and scored the NASA-TLX dimensions concerning the ILS approach after the go-around maneuver had been completed and the simulation had been stopped. The scale used in the scoring was 0 (= *low MWL*) to 100 (= *high MWL*).

Control Performance

The control performance of the participants was measured using ILS performance and climbing turn performance. ILS performance was evaluated between Points 3 and 4 (see Figure 2). ILS performance was determined based on the participants' horizontal and vertical errors with respect to the published ILS Z approach profile. In addition, the difference between the target airspeed and the airspeed maintained during the ILS approach was observed. ILS performance was determined after the flying task using mission playback. Ten measuring points were used to obtain the horizontal, vertical, and airspeed errors. The first nine measuring points were separated by 1 km. The distance between the ninth and the last measuring points was 0.9 km (see Figure 2). Averages for the horizontal, vertical, and airspeed errors were calculated for each participant and used as ILS performance measures. From the playback of the 3-min climbing turn maneuver, the participants' airspeed and bank angle errors during the maneuver were observed at every 15 s. Averages of airspeed and bank angle errors were calculated and used as performance measures for the turning climb maneuver.

Procedure

Before the trials, the participants were briefed about the Swing weighting method. They then weighted the workload contributions on the NASA-TLX dimensions with respect to a typical EFKU ILS Z33 instrument approach. They were also instructed how to report their MWL using the NASA-TLX workload form. The participants were given a normal mission briefing, after which they entered the FTD. The order in which the participants flew the in-flight and post-flight NASA-TLX missions was randomized.

Before the simulation was started, the participants rested in the FTD for 3 min. For the in-flight NASA-TLX mission, the FTD was initialized to Point 1 (see Figure 2). The pilots flew the ILS Z profile to DH, after which they executed a go-around and commenced a 3-min climbing turn maneuver. While in the climbing turn maneuver, the participants evaluated their MWL on the ILS approach and assigned the scores for the NASA-TLX dimensions. The participants referred to the pencil-and-paper version of the NASA-TLX form attached on their kneeboard and spoke aloud the scores for each dimension. The flight instructor copied the scores using the NASA-TLX paper and pencil version (available from: https://humansystems.arc.nasa.gov/groups/tlx/downloads/TLXScale.pdf). After 3 min of performing the climbing turn maneuver, the simulation was stopped.

The procedure for the post-flight NASA-TLX mission was identical to the in-flight NASA-TLX mission until the participants reached DH. In the post-flight mission, the participants did not evaluate their workload and provide NASA-TLX scores during the climbing turn maneuver. Instead, once the maneuver was completed, the simulation was stopped and a 3-min rest period was initiated. After the rest period, the participants assigned their NASA-TLX scores. As a result, the time delay in the post-flight mission was approximately 6 min.

Results

Data were analyzed using IBM SPSS statistics software (version 27). Table 1 summarizes the descriptive statistics for average altitude, airspeed, and heading errors for the ILS approaches during post-flight and in-flight NASA-TLX missions. For the purposes of calculating an estimate of the Type II error (β), the "true" best estimate of the pilots' control performance was defined as occurring in the post-flight NASA-TLX condition, when the participants were not under the additional workload imposed by completing the workload scale in flight. A t test did not reveal statistically significant differences in the mean altitude, $t(20) = -0.265$, $p = .793$; $\beta = 0.100$, airspeed, $t(20) = -0.057$, $p = .955$; $\beta = 0.062$, or heading, $t(20) = -0.458$, $p = .652$; $\beta = 0.125$), errors between the missions. The Type II error probabilities (β) in all cases were also small, suggesting that there was a low likelihood of the null hypothesis

Table 1. Means (*M*) and standard deviations (*SD*) of average altitude (m), airspeed (kn), and heading (degrees) errors during ILS approaches preceding post-flight and in-flight assignment of dimension scores ($N = 21$)

	Altitude		Airspeed		Heading	
	M	*SD*	*M*	*SD*	*M*	*SD*
Post-flight NASA-TLX ILS performance	9.17	4.50	9.10	3.69	0.43	0.36
In-flight NASA-TLX ILS performance	8.81	5.45	9.02	4.88	0.39	0.27

Note. ILS = instrument landing system.

Table 2. Means (*M*) and standard deviations (*SD*) of average bank angle (degrees) and airspeed (kn) errors during go-arounds with and without the additional task of score assignments ($N = 21$)

	Bank angle		Airspeed	
	M	*SD*	*M*	*SD*
Post-flight NASA-TLX go-around performance	2.03	1.30	5.28	2.67
In-flight NASA-TLX go-around performance	2.31	1.51	5.90	3.64

Table 3. Means (*M*) and standard deviations (*SD*) of weighted dimension scores and overall indices obtained in the in-flight NASA-TLX mission and the post-flight NASA-TLX mission ($N = 21$).

	MD		PD		TD		OP		EF		FR		OW	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
In-flight	10.52	3.05	4.44	4.50	8.59	3.40	6.76	4.01	9.84	3.44	4.27	3.36	44.42	21.75
Post-flight	10.48	3.26	4.35	4.34	8.37	3.68	6.25	3.19	8.83	3.29	3.76	2.94	42.05	20.70

Note. MD = mental demand; PD = physical demand; TD = temporal demand; OP = own performance; EF = effort; FR = frustration; OW = overall workload.

(the hypothesis of no difference) being false. In other words, it is quite probable that control performance in both missions was the same.

Table 2 presents the descriptive statistics of average bank angle and airspeed errors during go-arounds with and without the additional task of evaluating workload and completing NASA-TLX score assignments. A *t* test showed no statistically significant differences in the mean bank angle, $t(20) = -0.897$, $p = .380$; $\beta = 0.004$, and airspeed $t(20) = -1.232$, $p = .232$; $\beta = 0.003$, errors between the two types of go-arounds. The Type II error probabilities in both cases were also low, suggesting that control performance was the same in both NASA-TLX completion conditions. A Pearson product–moment correlation indicated strong positive correlations between the bank angle errors ($r = .498$, $p = .022$) and the airspeed errors ($r = .773$, $p = .000$) obtained from the go-arounds with and without in-flight NASA-TLX score assignments, further confirming this observation.

Table 3 summarizes the descriptive statistics of weighted dimension scores and overall MWL indices obtained during the in-flight and post-flight NASA-TXL missions. Table 4 presents *t*-test statistics and Type II error estimates for the scores and the workload indices. Table 5 describes the corresponding correlations for scores and indices.

The Type II error probabilities when comparing the weighted dimension scores and the overall MWL indices in the in-flight NASA-TLX mission and the post-flight NASA-TLX mission are all quite low, with the possible exception of the EF dimension (see Table 4). This largely indicates that workload values were the same in both missions. This is also supported by the correlation analyses reported in Table 5.

Table 4. Dependent *t*-test results and Type II error estimates of weighted dimension scores and overall indices obtained in the in-flight NASA-TLX mission and the post-flight NASA-TLX mission ($N = 21$)

	$t_{(20)}$	*p*	β
MD	0.052	.959	0.056
PD	0.334	.742	0.061
TD	0.408	.688	0.085
OP	0.650	.523	0.179
EF	1.560	.134	0.405
FR	1.037	.312	0.195
OW	0.939	.359	0.288

Note. MD = mental demand; PD = physical demand; TD = temporal demand; OP = own performance; EF = effort; FR = frustration; OW = overall workload.

Table 5. Pearson product–moment correlations (*r*) for weighted dimension scores and overall indices obtained in the in-flight NASA-TLX mission and the post-flight NASA-TLX mission ($N = 21$)

	MD	PD	TD	OP	EF	FR	OW
r	.506	.961	.757	.519	.611	.755	.259
p	.019	.000	.000	.016	.003	.000	0257

Note. MD = mental demand; PD = physical demand; TD = temporal demand; OP = own performance; EF = effort; FR = frustration; OW = overall workload.

Discussion

The task demand of the ILS approach in both flying missions was identical; thus, as would be expected, there was no significant difference in ILS control performance between the in-flight and post-flight NASA-TLX administration conditions (see Table 1). Altitude, heading, and airspeed errors were similar in both magnitude and variance. The in-flight evaluation of workload and administration of NASA-TLX occurred during the climbing turn maneuver, which followed the ILS approach and the go-around. If the administration of NASA-TLX contributed a significant information processing overhead (see Harris, 2011), it would be expected that there would be a significant difference in bank angle and airspeed errors during the climbing turn. However, there was no significant difference in climbing turn control performance between the two task demand conditions (see Table 2) and the magnitude of the Type II errors suggested that the performance was likely to be the same. This implies that the cognitive overhead of completing the NASA-TLX scoring was low, and that the administration of NASA-TXL did not intrude on the task performance. In this study, the participants referred to a printed version of the NASA-TLX form and spoke aloud the scores. In most aircraft, the intercom is recorded and the scores can be obtained from the cockpit recording. As such, there is no need to manually complete the NASA-TLX form during flight.

In comparing the individual weighted dimension scores and the overall MWL indices when NASA-TLX was undertaken in-flight and post-flight, there were also no significant differences in either scores or overall indices (see Tables 3 and 4). Moreover, the magnitude of the Type II errors suggested that there was a high likelihood that the scores and the indices were substantially the same. This implies that in this case the workload estimation and scoring task did not interfere with the climbing turn task, imposing more workload. This is contrary to the suggestions of both Corwin (1992) and Annett (2002b). Moroney et al. (1995) also observed a time error when delaying the making of workload scorings. However, they observed that the effect increased as the delay became longer and was negligible with up to 30-min delay. In this study, the delay between scorings attributable to the go-around task was around 6 min, which may have helped in this respect.

The results from this study are unusual in that the nonsignificant results, when considered in conjunction with the corresponding Type II error estimates, provide confidence that the data from the two experimental conditions are substantially the same. The correlation coefficients (see Table 5) give an indication of the variance shared between the two conditions and indicate that a rise in one dimension score may be concomitant with a rise in the other dimension score. The correlation coefficient, however, does not suggest that the means or the variances of the two scores are equivalent in magnitude. Hence, the correlation coefficients in Table 5 should be interpreted in conjunction with the contents of Table 4 (*t*-test and Type II error data). From a methodological perspective, this implies both that the in-flight collection of NASA-TLX data is feasible when the flight segment during which NASA-TLX is administered is not too demanding and that post-flight data collection provides the same workload estimates – at least if the data collection is not greatly delayed. The correlation coefficients for the individual dimensions when using the Swing weighting method (Von Winterfeldt & Edwards, 1986) are of the same order as those reported by Corwin (1989), who utilized the original NASA-TLX weighting procedure.

NASA-TLX has been shown to have methodological challenges regarding the impact of time delay when the scoring of dimensions is carried out (Annett, 2002b) and the way the dimension weights are assigned (Virtanen et al., 2021). The results of this study reveal that the issue related to the time delay between the assignment of scores can be overcome by conducting the scoring during the simulated flying mission. However, future research is needed to evaluate how the in-flight scoring procedure affects control accuracy and NASA-TLX scores – especially if it is conducted during a real flight. Such trials should encompass scenarios of varying degrees of difficulty and in particular more demanding phases of flight. While the results of this paper are encouraging, it is still recommended to use multiple measures when assessing MWL as it minimizes the potential biases of individual measures.

Conclusion

Overall, NASA-TLX seems to be a reliable near-real-time technique for estimating pilots' subjectively experienced MWL – at least if it is used in an experimental context.

References

Aasman, J., Mulder, G., & Mulder, L. J. (1987). Operator effort and the measurement of heart-rate variability. *Human Factors, 29*(2), 161–170. https://doi.org/10.1177/001872088702900204

Alaimo, A., Esposito, A., & Orlando, C. (2021). Pilot's mental workload nonlinear correlation among objective and subjective measurements. *Journal of Physics Conference Series, 1786*(1), 1–5. https://doi.org/10.1088/1742-6596/1786/1/012011

Annett, J. (2002). Subjective rating scales in ergonomics: A reply. *Ergonomics, 45*(14), 1042–1046. https://doi.org/10.1080/00140130210166762

Annett, J. (2002). Subjective rating scales: Science or art? *Ergonomics, 45*(14), 966–987. https://doi.org/10.1080/00140130210166951

Banks, V., Plant, K., & Stanton, N. (2018). Development of a usability evaluation framework for the flight deck. In R. Charles & J. Wilkinson (Eds.), *Contemporary ergonomics and human factors* (pp. 1–7). CIEHF.

Bishop, A., MacNeil, E., & Izzetoglu, K. (2021). Cognitive workload quantified by physiological sensors in realistic immersive settings. In D. Schmorrow & C. Fidopiastis (Eds.), *Augmented cognition* (pp. 119–133). Springer.

Casali, J., & Wierwille, W. (1984). On the measurement of pilot perceptual workload: A comparison of assessment techniques addressing sensitivity and intrusion issues. *Ergonomics, 27*(10), 1033–1050. https://doi.org/10.1080/00140138408963584

Corwin, W. (1989). *The NASA task load index as a measure of pilot workload in the commercial transport environment* (Tech. Rep. No. 892382). SAE International.

Corwin, W. (1992). In-flight and postflight assessment of pilot workload in commercial transport aircraft using the subjective workload assessment technique. *The International Journal of Aviation Psychology, 2*(2), 77–93. https://doi.org/10.1207/s15327108ijap0202_1

Efthymiou, M., Fichert, F., & Lantzsch, O. (2019). Workload perception of air traffic control officers and pilots during continuous descent operations approach procedures. *Aviation Psychology and Applied Human Factors, 9*(1), 2–11. https://doi.org/10.1027/2192-0923/a000154

Eggemeier, F. T., Crabtree, M. S., & LaPointe, P. A. (1983, October). The effect of delayed report on subjective ratings of mental workload. *Proceedings of the Human Factors Society Annual Meeting, 27*(2), 139–143. https://doi.org/10.1177/154193128302700205

Etherington, T., Kramer, L., Bailey, R., Kennedy, K., & Stephens, C. (2016). *Quantifying pilot contribution to flight safety for normal and non-normal airline operations, 2016 IEEE/AIAA 35th Digital avionics systems conference* (pp. 1–14). IEEE.

Harris, D. (1991). The importance of the Type II error in aviation safety research. In E. Farmer (Ed.), *Stress and error in aviation* (pp. 151–157). Aldershot.

Harris, D. (2011). *Human performance on the flight deck*. Ashgate Publishing.

Hart, S. (2006). NASA-task load index (NASA-TLX): 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. Hanckock & N. Meshkati (Eds.), *Human mental workload* (pp. 139–183). North-Holland.

Jorna, P. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics, 36*(9), 1043–1054. https://doi.org/10.1080/00140139308967976

Li, F., & Chen, K. (2020). High-level review principles for human-machine interface design of civil aircraft flight deck. In R. Stanton (Ed.), *Advances in Human Aspects of Transportation: Proceedings of the AHFE 2020 Virtual Conference on Human Aspects of Transportation, July 16-20, 2020, USA* (pp. 624–631). Springer.

Mansikka, H., Simola, P., Virtanen, K., Harris, D., & Oksama, L. (2016). Fighter pilots' heart rate, heart rate variation and performance during instrument approaches. *Ergonomics, 59*(10), 1344–1352. https://doi.org/10.1080/00140139.2015.1136699

Mansikka, H., Virtanen, K., & Harris, D. (2018). Comparison of NASA-TLX scale, Modified Cooper-Harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks. *Ergonomics, 62*(2), 246–254. https://doi.org/10.1080/00140139.2018.1471159

Mansikka, H., Virtanen, K., & Harris, D. (2019). Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft. *IEEE Transactions on Human-Machine Systems, 49*(1), 1–9. https://doi.org/10.1109/THMS.2018.2874186

Matthews, G., Reinerman-Jones, L., Wohleber, R., Lin, J., Mercado, J., & Abich, J. (2015). Workload is multidimensional, not unitary: What now? In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of Augmented Cognition* (pp. 44–55). Springer.

Moroney, W., Biers, D., & Eggemeier, F. (1995). Some measurement and methodological considerations in the application of subjective workload measurement techniques. *The International Journal of Aviation Psychology, 5*(1), 87–106. https://doi.org/10.1207/s15327108ijap0501_6

Muth, E., Moss, J., Rosopa, P., Salley, J., & Walker, A. (2012). Respiratory sinus arrhythmia as a measure of cognitive workload. *International Journal of Psychophysiology, 83*(1), 96–101. https://doi.org/10.1016/j.ijpsycho.2011.10.011

Paas, F., & Van Merriënboer, J. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors, 35*(4), 737–743. https://doi.org/10.1177/001872089303500412

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

Silva, M., Macêdo, T., Lourenço, C., Rehder, I., Marchiori, A., Cesare, M., Corters, R., Cardoso, M. Jr., & Villani, E. (2021). *Mental workload assessment in military pilots using flight simulators and physiological sensors. International symposium on human mental workload: Models and applications* (pp. 99–115). Springer.

Sohn, S., & Jo, Y. (2003). A study on the student pilot's mental workload due to personality types of both instructor and student. *Ergonomics, 46*(15), 1566–1577. https://doi.org/10.1080/0014013031000121633

Tsang, P., & Vidulich, M. (1994). The roles of immediacy and redundancy relative subjective workload assessment. *Human Factors, 36*(3), 503–513. https://doi.org/10.1177/001872089403600307

Vidulich, M. A. (1989). The use of judgment matrices in subjective workload assessment: The Subjective Workload Dominance (SWORD) Technique. *Proceedings of the Human Factors Society Annual Meeting, 33*(20), 1406–1410. https://doi.org/10.1177/154193128903302009

Virtanen, K., Mansikka, H., Kontio, H., & Harris, D. (2021). Weight watchers: NASA-TLX weights revisited. *Theoretical Issues in Ergonomics Science, 23*(6), 725–748. https://doi.org/10.1080/1463922X.2021.2000667

Von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioural research*. Cambridge University Press.

Wilson, G. (1993). Air-to-ground training missions a psychophysiological workload analysis. *Ergonomics, 36*(9), 1071–1087. https://doi.org/10.1080/00140139308967979

Young, M., & Stanton, N. (2002). It's all relative: Defining mental workload in the light of Annett's paper. *Ergonomics, 45*(14), 1018–1020. https://doi.org/10.1080/00140130210166816

History
Received November 14, 2023
Revision received May 27, 2024
Accepted June 6, 2024
Published online July 26, 2024

Conflict of Interest
No potential conflict of interest was reported by the authors.

Publication Ethics

Informed consent was obtained from all participants included in the study. According to the guidelines of the Finnish National Board on Research Integrity, the research configuration of this paper was such that it did not require an ethical review statement from a human sciences ethics committee.

ORCID

Heikki Mansikka
https://orcid.org/0000-0002-7261-6408
Don Harris
https://orcid.org/0000-0002-2113-8848

Heikki Mansikka

Systems Analysis Laboratory
Aalto University
P.O. Box 11110
00076 Aalto
Finland
heikki.mansikka@aalto.fi

Heikki Mansikka (PhD) is an adjunct professor at the Department of Military Technology, National Defence University, Finland. He is a former F/A-18 fighter pilot and his research interests include human and team performance in air combat and other complex systems.

Don Harris (PhD) is Professor of Human Factors at Coventry University, UK. He is a fellow of the Institute of Ergonomics and Human Factors and a chartered psychologist. Don Harris is heavily involved in the design and development of the next generation of advanced pilot interfaces and the certification of civil flight decks.

Kai Virtanen (PhD) is a professor in the joint professorship of Operations Research With Systems Analysis Laboratory, Department of Mathematics and Systems Analysis, Aalto University, Finland, and the Department of Military Technology, National Defence University, Finland. His research interests include decision analysis, dynamic optimization, simulation optimization, and human performance in complex systems.