Comparison of NASA-TLX scale, Modified Cooper-Harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks

Mansikka, H, Virtanen, K & Harris, D

DOI 10.1080/00140139.2018.1471159
ISSN 0014-0139
ESSN 1366-5847
Publisher: Taylor and Francis

Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: http://www.tandfonline.com/loi/terg20

Comparison of NASA-TLX scale, Modified Cooper-Harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks

Heikki Mansikka, Kai Virtanen & Don Harris

To cite this article: Heikki Mansikka, Kai Virtanen & Don Harris (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, DOI: 10.1080/00140139.2018.1471159

To link to this article: https://doi.org/10.1080/00140139.2018.1471159

Publisher: Taylor & Francis

Journal: *Ergonomics*

DOI: http://doi.org/10.1080/00140139.2018.1471159

Comparison of NASA-TLX scale, Modified Cooper-Harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks

Heikki Mansikka (corresponding author)

Nil

P.O.Box 35566 Abu Dhabi, UAE. heikki.mansikka@gmail.com

Heikki Mansikka is a retired Air Force fighter pilot (LtCol). He has a PhD from Coventry University, Aviation Human Factors and a MA from Kings' College London, Defence Studies. He is currently a Senior Manager, Pilot Performance in Etihad Airways, Abu Dhabi, UAE. His research concentrates on pilot performance.

Kai Virtanen

Department of Mathematics and Systems Analysis, Aalto University, Helsinki, Finland

Kai Virtanen is a professor of Operations Research at Department of Mathematics and Systems Analysis, Aalto University, Finland and at Department of Military Technology, National Defence University of Finland. His research interests include decision analysis, dynamic optimization, simulation-optimization and human performance in complex systems.

Don Harris

Faculty of Engineering and Computing, Coventry University, Coventry, United Kingdom

Don Harris is Professor of Human Factors at Coventry University. He is a Fellow of the Institute of Ergonomics and Human Factors and a Chartered Psychologist. Don is heavily involved in the design and development of the next generation of advanced pilot interfaces and the certification of civil flight decks

Abstract: The sensitivity of NASA-TLX scale, modified Cooper-Harper (MCH) scale and the mean inter-beat interval (IBI) of successive heart beats, as measures of pilot mental workload (MWL), were evaluated in a flight training device (FTD). Operational F/A-18C pilots flew instrument approaches with varying task loads. Pilots' performance, subjective MWL ratings and IBI were measured. Based on the pilots' performance, three performance categories were formed; high-, medium- and low-performance. Values of the subjective rating scales and IBI were compared between categories. It was found that all measures were able to differentiate most task conditions and there was a strong, positive correlation between NASA-TLX and MCH scale. An explicit link between IBI, NASA-TLX, MCH and performance was demonstrated. While NASA-TLX, MCH and IBI have all been previously used to measure MWL, this study is the first one to investigate their association in a modern FTD, using a realistic flying mission and operational pilots.

Keywords: pilot mental workload; NASA-TLX scale; modified Cooper-Harper scale; inter-beat interval

Practitioner summary: NASA-TLX scale, Modified Cooper-Harper (MCH) scale and the inter-beat interval (IBI) were evaluated in a flight training device. All measures were able to differentiate most task conditions and there was a positive correlation between NASA-TLX and MCH scale. An explicit link between IBI, NASA-TLX, MCH and performance was demonstrated.

1. Introduction

Pilot mental workload (MWL) characterises the demands imposed by tasks on limited mental resources when a desired level of performance is to be maintained (Wickens 2008; Vicente, Thornton, and Moray 1987). Based on this rather common definition, the expenditure of the mental resources is considered to vary for three reasons. First, variations in the task demand cause variations in the amount of mental resources required to satisfy the demand (Wickens 1991). Second, the available mental resources define the portion of the overall mental resources required to achieve a desired level of performance (Norman and Bobrow 1975). Third, the level of desired or acceptable performance dictates the amount of the voluntary mental resource investment or effort (Hockey 1997). For example, performance variations between two fighter pilots conducting the same flying task may result from unequal cognitive resources or different levels of effort. Likewise, if the same pilots generate equal levels of performance, they may need to invest different levels of effort and may have to expend different proportions of their mental resources. When the task demands are kept similar for both pilots and they are equally willing to invest effort on the tasks, the resulting performance differences result from differences in their information processing capacity (Mulder 1986; Kahneman 1973). However, when pilots are exposed to high or extreme task demands, some will deplete their mental resources sooner than others. Once there is no more mental capacity left to compensate for the increasing task demand, performance will begin to degrade regardless of the level of effort applied (Hockey

1997; Williges and Wierwille 1979). When the performance degradation becomes significant, flight safety and operational effectiveness are at risk of being compromised. When new aircraft systems are designed or new tactics, techniques and procedures (TTPs) are developed, it is necessary to evaluate both the human-machine performance and MWL. Measuring task performance without evaluating MWL provides an incomplete understanding of the effectiveness and the safety margin of a human-machine system or TTP.

A variety of measures are available to assess MWL. Most empirical measures can be categorised either as behavioral, subjective or physiological. Not all measures are applicable for all purposes. When the different techniques for MWL measurement are considered for a particular application, five major criteria should be considered; sensitivity, diagnosticity, intrusiveness, implementation requirements and operator acceptance (for a more detailed discussion, see Wierwille and Eggemeier, [1993]; O'Donnell, Eggemeier, and Thomas [1986]). The different characteristics of the measures make some of them more suitable for the fighter aviation environment than others.

The subjective MWL measures utilise the operator's subjectively experienced MWL, i.e. how a person feels when doing a task (Johanssen et al. 1979). The non-intrusiveness, ease of use and low-cost implementation of the subjective MWL measures are some of the features that motivate their usage (see e.g., Fallahi et al., [2016]; Gabriel, Ramallo, and Cervantes [2016], Akyeampong [2014]; Prichard., Bizo, and Stratford., [2011]; DiDomenico and Nussbaum [2008]; Newell and Mansfield [2008]; Sato et al., [1999]) For subjective MWL measuring, there are many different methods to choose from. These include, but are not limited to, methods such as NASA Task Load index (NASA-TLX) (Hart and Staveland 1988) and Modified Cooper-

Harper scale (MCH) (Wierwille and Casali 1983), which are both widely used in an aviation domain (Zhang et al. 2009; Di Nocera, Camilli, and Terenzi 2007; Dennis and Harris 1998; van Westrenen 1996; Moroney et al. 1992; Hill et al. 1992; Battiste and Bortolussi 1988; Skipper, Rieger, and Wierwille, 1986; Casali and Wierwille 1984; Casali and Wierwille 1983).

Some techniques, like NASA-TLX, are multidimensional measuring scales, which use several different dimensions to assess MWL. NASA-TLX consists of six subscales that represent different dimensions of MWL: mental demand, physical demand, temporal demand, frustration, effort, and performance (Hart and Staveland 1988). The unidimensional scales, such as MCH, utilize just one dimension and their overall sensitivity are sometimes questioned as they ignore the multidimensional nature of the human information processing and don't even attempt to distinguish the task processing demands on different cognitive modalities or stages (Wickens 2008; Hill et al. 1992). The basic assumption of the subjective MWL measures is that if a person experiences high MWL, stress or frustration, then s/he has high MWL, stress or frustration – regardless of the indications of the other measures.

Variations in arousal, effort and general activation level cause physiological changes. This has motivated the use of various physiological measures as indices of MWL. The major advantage of physiological measures is their ability to provide continuous, real time monitoring of the operator state (Jorna 1993). Another advantage is their objectivity, which also increases their utility in scenarios where it is reasonable to expect that subjective opinions are not accurate (Gopher and Donchin 1986). Although MWL cannot be measured directly, the heart's responses to the neurological modulation provide an indirect method for its measurement; nerve activity causes electronic impulse transmissions in and around the heart, which can be recorded and

interpreted with an electrocardiograph (ECG). A normal ECG consists of a P-wave, a QRS complex, followed by a T-wave and U-wave, each representing different de- and repolarization phases within the heart's muscular cells. Once the QRS complexes are detected from the ECG, the normal-to-normal (NN) inter-beat interval and differences between the NN intervals can be determined. When NN intervals are analysed, a decreased NN interval or a lowered mean of the inter-beat intervals between successive heart beats (IBI) can be used as indirect indicators of increased MWL. IBI and IBI interval differences have been successfully used to measure task demand variations both in a flight simulator and in actual flight (Mansikka et al. 2016a, 2016b; Dahlstrom and Nahlinder 2009; Svensson and Wilson 2002; Svensson, Angelborg-Thanderz, and Wilson 1999; Veltman and Gaillard 1998; Svensson et al. 1997; Ylönen et al. 1997; Wilson 1993; Roscoe 1993; Jorna 1993; Aasman, Mulder, and Mulder 1987; Roscoe 1975). It is a common practice to use the R-wave peak as a reference point in measurements as it is typically the strongest wave and can therefore be easily detected even in noisy conditions. To emphasize the reference point used, the literature typically uses terms R-wave to R-wave (RR) interval and RR interval difference (or IBI and IBI difference) (ChuDuc, NguyenPhan, and NguyenViet 2013; Opmeer 1973). Out of the different methods available, this study used time domain methods to analyse RR intervals and RR interval differences. The time domain analysis techniques are based on the statistical analysis of the series of successive RR intervals. In its simplest form, the statistical analysis is used to determine IBI. It was expected that the simplicity of the method would encourage a broader audience to utilise the methods used in this study.

Unfortunately, while the physiological measures allow objective, continuous and real time monitoring of the pilot's state, their sensitivity may become limited if utilised in real flying missions. The bodily interactions and physiological responses to external

stimuli may make it difficult or impossible to use the physiological measures of MWL during real flight. For example, pupillary diameter may be affected not only by variations in the information processing demands, but also by the variations in the eye's fixation distance or ambient lightning. In a similar fashion, cardiological responses can be affected by the blood pressure variations, body temperature and arterial pressure. In fighter aviation, factors like extreme cockpit temperatures, exposure to direct sunlight and high G-loads can generate physiological responses which can, if not properly controlled, be falsely interpreted as MWL responses. Subjective measures, on the other hand, are easy to employ in simulated and real fighter missions, and they generally enjoy high face validity and broad operator acceptance. These measures, however, have been criticised for their inherent tendency to generate time error as the data collection is typically conducted after the activity, and for the fact that the subjects must compare and arrange their past sensations to a rating scale (Annett 2002). In addition, the number of different task features and the phasing of high and low task demand events can affect the subjective perception of workload (Wierwille, Rahimi, and Casali 1985). The accuracy of subjective measures is therefore reliant of the participants' ability to memorize their perceived workload in retrospect. The nature of the subjective measures' scales has also been criticised. While the operator's inner sensations may be represented on a numerical scale, the scale itself is ordinal. All subjective workload measuring scales lack the interval and ratio properties and there are no universal units for the scales. That is, even if the subjective feelings or sensations are given numeric values, the distance between the values may not be equal (Casner and Gore 2010; Annett 2002; Michell 1997). More importantly, the unreliability of their main instrument – the human operator – has been questioned. The potential operator bias is especially problematic if the amount of MWL is used as a personnel evaluation or

selection criteria. If that is the case, it is possible that while a person reports acceptable MWL, stress or frustration, s/he may still experience excessive MWL, stress or frustration. In such a context, the operator is a highly unreliable and insensitive measuring instrument (Gopher and Donchin 1986). In summary, both subjective and physiological MWL measures may be effective and sensitive in certain situations and highly ineffective or unreliable in others. Luckily, when MWL is measured in a flight training device (FTD), most of the potentially disruptive external stimuli discussed above can be ruled out. In addition, the non-punitive context used in this study greatly reduces the risk of potential pilot biases when subjective MWL measures are used.

This study investigated if both subjective and physiological measures are sensitive to varying task demands when fighter pilots' MWL is measured during a simulated flying mission. Should this be the case, system designers and TTP developers would have more MWL measures to choose from and could select the most appropriate one for each test setting. Also, if the different measures of MWL provide similar results, it enables the utilisation of multiple MWL measures. This is essential as the complexity of the human information processing system may require multiple MWL measures to be used to reveal the demands of a single human-machine system (Wickens 2008). Furthermore, the use of multiple measures allows them to be cross validated.

As discussed by Carayon et al. (2015, 2006), Dekker (2012) and Verwey and Veltman (1996), complex human-machine systems should be assessed in their context. The complexity and uniqueness of many human-machine systems make it difficult to generalise the task demands of one system to other systems. While NASA-TLX, MCH and IBI have all been previously used to measure MWL, this study is the first one to investigate their association in a modern FTD, using a realistic flying mission and operational fighter pilots. Hence, the objective of this study was to investigate if there

are differences between NASA-TLX, MCH and IBI, as measures of MWL, when the fighter pilots' MWL is measured during the simulated flying mission.

2. Method

2.1 Participants

Workload and performance data were collected from Finnish Air Force McDonnell-Douglas F/A-18C pilots. Limited by the available FTD time to run the trials and the volunteering pilots' operational assignments at the time of the study, the subsequent number of participants was 27. The participants' qualifications and experience levels on the aircraft type varied from a wingman to an experienced instructor pilot. As a result, the participants' average flying experience on the aircraft type was 627 flight hours (SD= 476). Participants were declared medically fit to fly. Relevant nutrition and activity data was collected from each participant for the 12 hours prior to participating. A written, informed consent was obtained from each participant.

2.2 Test Design

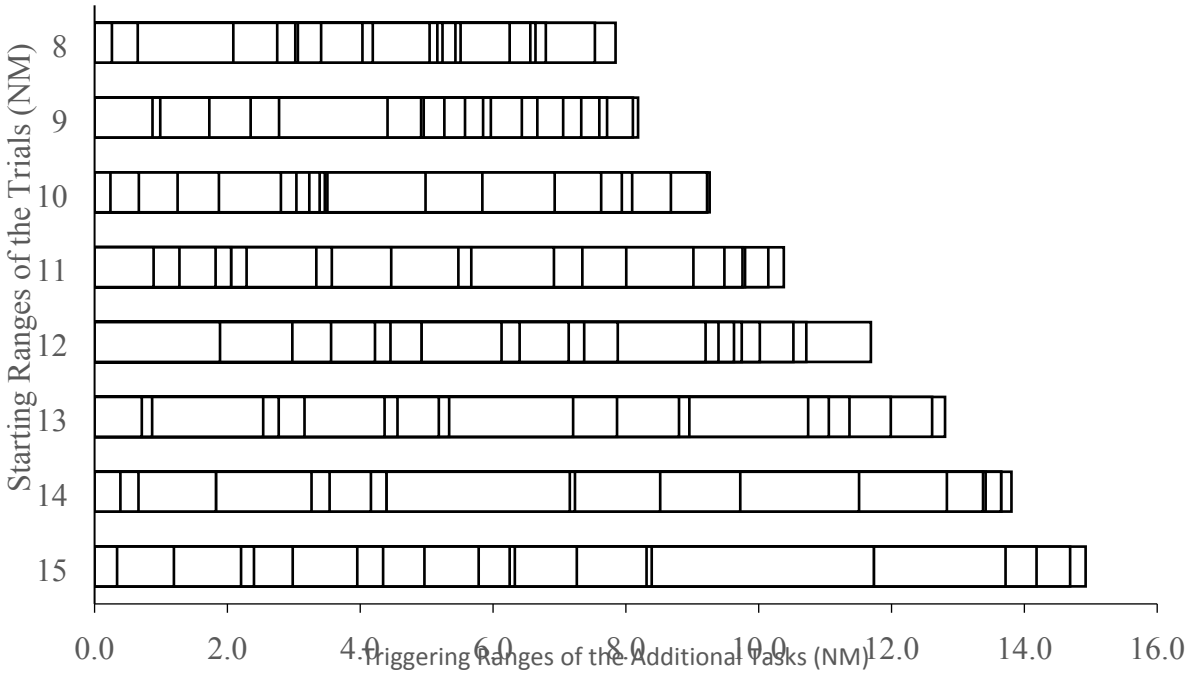
A weapon tactics and situational awareness trainer (WTSAT) was used for the flight task. WTSAT is a non-motion FTD with a 135 degree field of view and a fully functional cockpit, which allow it to be used for basic and advanced flight training. For this study, a cross wind of 80 degrees, 10 knots (5.14 m/s) with moderate gusts was set. Full instrument meteorological conditions were set at and above 200 ft (60 m). Below 200 ft (60 m) the meteorological conditions were set in order to satisfy the minimum meteorological conditions required for a visual landing.

The flying task consisted of number of instrument landing system (ILS) approaches. All participants were highly familiar with the flying task as they had been

routinely flying ILS approaches both with F/A-18 and with their earlier aircraft types. ILS is a precision approach procedure, which provides a pilot with both horizontal and vertical control cues throughout the approach profile. The participants' task was to fly the approach profile with minimal horizontal and vertical errors while completing normal, self-paced approach and landing preparations, e.g., using the radios, configuring the aircraft for landing and cross checking the flight instruments. In addition to the self-paced tasks, set of additional tasks were triggered in each trial at predefined, randomised ranges. These tasks included ten warnings or cautions requiring immediate pilot action, seven radio calls requiring pilot acknowledgement and three requests to change frequencies or altimeter settings. Figure 1 summarises the triggering ranges of the additional tasks.

For each trial, the simulated aircraft was initially set to 2,000 ft (607 m) above ground level. The participants were briefed to maintain a constant airspeed throughout the trials. The temporal demand of the flying task was manipulated by varying the trials' starting ranges. The starting ranges varied from 8 NM (14.8 km) to 15 NM (27.8 km) with 1 NM (1.9 km) increments. As a result, each participant flew eight trials where the time available for the trials varied from 6 minutes 22 seconds to 3 minutes 27 seconds with each 1 NM (1.9 km) decrement in starting range reducing the time available by 25.5 seconds. The order of the starting ranges was randomised between subjects. Each trial ended at the ILS specific decision height of 200 ft (60 m). The trials were separated by a rest period lasting approximately three minutes.

Figure 1. Triggering ranges of the additional tasks for each starting range of the trials. The vertical lines in each trials' range bar indicate the additional tasks' triggering ranges within that trial. Each range bar has 20 additional tasks.



2.3 Procedure

Like some recent studies (Mansikka et al. 2016a; Wang et al. 2016) the time available for each trial, or time pressure, was used as an independent variable in the test design. The ILS performance score, mean IBI value, NASA-TLX score and MCH score from each trial were used as dependent variables.

The ILS performance was rated using an existing instrument proficiency test rating scale. The ILS performance rating was based on three components: the horizontal and vertical deviations from the optimal ILS approach profile and the deviations from the target airspeed. During the ILS profile, these deviations were recorded and scored independently at every 0.5 NM (0.9 km). The possible values of each ILS performance score component ranged from 5 (best performance) to 0 (worst performance). The mean of the component scores was used as an ILS performance score. The ILS performance scoring was undertaken by a qualified examiner pilot.

ECG was recorded using Mind Media NeXus-10 MKII system. Three electrodes

were placed below the left (negative) and right (ground) clavicle and the left costal cartilage (positive), respectively. Three minute samples were taken from each trial. In addition, a three minute pre-test rest sample was taken from each participant. ECG data was first analysed with Biotrace+ software (version V2012C) from where the samples were exported to Kubios HRV 2.2 software for further analysis and artefact removal. All artefacts were detected and removed manually and noisy data were excluded from further analysis. ECG measuring, manipulation and interpretation were done in accordance with the guidance in Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (Camm et al. 1996).

Before the trials, each participant was trained to use NASA-TLX and MCH scales. After each trial, the participants gave NASA-TLX and MCH scores for the trial. For NASA-TLX, the scores of the scale's six dimensions (or subscales) were summed and used as a unidimensional NASA-TLX score. Raw NASA-TLX scores were used, i.e. the scores were not weighted. The use of the raw NASA-TLX scores is a common modification of the original scale and makes it somewhat easier to use (Hart 2006, Byers et al. 1989).

3. Results

3.1 Treatment of Data

The ILS performance scores were used to create three performance categories. A high-performance category was formed by selecting each participant's trial with the highest ILS performance score. A medium-performance category was generated by selecting each participant's trial that had the average ILS performance score when compared to each participant's other trials. Finally, a low-performance category was formed by selecting each participant's trial with the lowest ILS performance score. The different

performance categories were used as measurement points for MWL by retrieving mean IBI values as well as NASA-TLX and MCH scores for each trial in the three performance categories. In addition, mean IBI values for the pre-test rests were retrieved

The normality of the values of performance, IBI, MCH, NASA-TLX and NASA-TLX subscales in each performance category was tested using both the graphical methods and Saphiro-Wilk test. Once the approximate normality requirement was confirmed, the data were analysed using repeated measures ANOVA. Violation of sphericity was handled with a Greenhouse-Geisser correction. Pairwise comparisons were carried out using the paired t-test. Finally, Pearson product-moment correlation was run to determine the association between NASA-TLX and MCH in the different performance categories.

3.2 Analysis

The high-performance category had the mean ILS score of 4.73 (SD=0.11). For the medium-performance category the mean ILS score was 3.72 (SD=0.36), and for the low-performance category the mean ILS score was 2.43 (SD=0.84). The mean ILS performance scores between all three performance categories were significantly different ($F(1.243,32.308)=161.382$, $p<0.001$, partial $\eta 2=0.861$). A significant difference between the three performance categories was found for MCH mean values ($F(1.918,49.862)=72.938$, $p<0.001$, partial $\eta 2=0.737$) and for NASA-TLX mean values ($F(1.962,51.024)=93.468$, $p<0.001$, partial $\eta 2=0.782$). Also, IBI mean values were significantly different between the three performance categories and pre-test rest ($F(2.078, 54.033)=39.130$, $p<0.001$, partial $\eta 2=0.601$). Table 1 presents the descriptive statistics for IBI, NASA-TLX and MCH across the three measurement points. Table 1 also shows the descriptive statistics for IBI pre-test rest values.

Table 1. Means and standard deviations (SD) for IBI, NASA-TLX and MCH at the measurement points (N=27).

		Rest		High-Performance Category		Medium-Performance Category		Low-Performance Category	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
IBI	(ms)	851.80	175.52	730.17	157.56	682.67	121.78	678.20	127.63
NASA-TLX	(-)	-	-	14.93	6.42	33.48	9.36	39.04	7.86
MCH	(-)	-	-	2.44	1.12	5.37	1.57	6.67	1.92

The results of the pairwise comparisons between the performance categories are summarised in Table 2. Both NASA-TLX and MCH scores showed significant differences between the three performance categories ($p<0.01$). Mean IBI values showed significant difference between the high-performance category and the medium-performance category, as well as between the high-performance category and the low-performance category ($p<0.05$). There was no significant IBI value difference between the low-performance category and the medium-performance category ($p>0.05$). In addition, the paired t-tests revealed significant differences between the mean rest IBI values and mean IBI values of all three performance categories ($p<0.001$).

Table 2. Values of IBI, MCH and NASA-TLX test statistics and pairwise comparisons between the measurement points (N=27).

		High-performance category - Medium performance category			High- performance category - Low performance category			Medium- performance category - Low-performance category		
		Mean Diff	Std. Error	t-value	Mean Diff	Std. Error	t-value	Mean Diff	Std. Error	t-value
IBI	(ms)	47.5	13.3	3.6**	52.0	19.6	2.7*	4.5	13.1	0.3
NASA-TLX	(-)	-18.6	2.0	-9.4***	-24.1	1.8	-13.3***	-5.6	1.8	-3.2**
MCH	(-)	-3.0	0.3	-8.7***	-4.2	0.4	-10.7***	-1.3	0.3	-3.8**

***p<0.001; **p<0.01; *p<0.05 (N=27)

NASA-TLX subscale values were further analysed. The values within each subscale increased as the ILS performance scores decreased. Table 3 summarises the

descriptive statistics for NASA-TLX subscale values across the measurement points.

Table 3. Means and standard deviations (SD) for NASA-TLX subscale values at the measurement points (N=27).

	High-Performance Category		Medium-Performance Category		Low-Performance Category	
	Mean	SD	Mean	SD	Mean	SD
Mental Demand	2.37	1.28	6.33	1.90	7.11	1.60
Physical Demand	1.93	1.14	5.70	2.03	6.26	1.70
Temporal Demand	1.81	1.11	6.00	2.24	7.19	1.62
Effort	3.19	1.90	6.81	2.00	7.52	1.81
Performance	3.30	1.17	4.81	1.80	6.37	2.19
Frustration	2.33	1.24	3.70	2.05	4.63	1.90

Repeated measures ANOVA revealed that the subscale values of NASA-TLX differed significantly between the measurement points: Mental Demand, $F(1.937,50.369)=106.849$, $p<0.001$, partial $\eta 2=0.804$; Physical Demand, $F(1.925,50.051)=88.927$, $p<0.001$, partial $\eta 2=0.774$; Temporal Demand, $F(1.824,47.415)=87.213$, $p<0.001$, partial $\eta 2=0.770$; Effort, $F(1.898,49.339)=58.416$, $p<0.001$, partial $\eta 2=0.692$; Performance, $F(1.734,45.093)=23.075$, $p<0.001$, partial $\eta 2=0.470$; Frustration, $F(1.998,51.938)=13.081$, $p<0.001$, partial $\eta 2=0.335$. Furthermore, pairwise comparisons revealed that the scores of all NASA-TLX subscales differed significantly between the high- and low-performance categories, as well as between the high- and medium-performance categories. However, only the 'Mental Demand', 'Temporal Demand' and 'Performance' values changed significantly between the medium- and low-performance categories. Table 4 summarises NASA-TLX subscales' pairwise comparisons between the performance categories.

Table 4. Values of NASA-TLX subscale test statistics and pairwise comparisons between the measurement points (N=27).

	High-performance category - Medium-performance category			High-performance category - Low performance category			Medium-performance category - Low-performance category		
	Mean Diff	Std. Error	t-value	Mean Diff	Std. Error	t-value	Mean Diff	Std. Error	t-value
Mental Demand	-4.0	0.4	-10.6***	-4.7	0.3	-13.8***	-0.8	0.3	-2.4*
Physical Demand	-3.8	0.4	-9.8***	-4.3	0.3	-13.2***	-0.6	0.3	-1.6
Temporal Demand	-4.2	0.5	-9.1***	-5.4	0.4	-15.1***	-1.2	0.5	-2.6*
Effort	-3.6	0.4	-9.1***	-4.3	0.5	-9.1***	-0.7	0.4	-1.7
Performance	-1.5	0.4	-4.0***	-3.1	0.4	-7.0***	-1.6	0.5	-2.9**
Frustration	-1.4	0.4	-3.1**	-2.3	0.5	-5.0***	-0.9	0.5	-2.0

***p<0.001; **p<0.01; *p<0.05 (N=27)

Pearson product-moment correlation was run to determine the relationship between NASA-TLX and MCH scores in the different performance categories. There was a significant, positive correlation between NASA-TLX and MCH scores in the high-performance category ($r=0.897$, $p<0.001$, $N=27$), the medium-performance category ($r=0.654$, $p<0.001$, $N=27$) and the low-performance category ($r=0.821$, $p<0.001$, $N=27$). NASA-TLX and MCH scores did not correlate with mean IBI values ($p>0.05$).

4. Discussion

The objective of this study was to investigate if there are differences between NASA-TLX, MCH and IBI when the fighter pilots' MWL was measured during the simulated flying mission. NASA-TLX, MCH and IBI have all been successfully and widely used to (indirectly) measure MWL (Mansikka et al. 2016b; Wang et al. 2016; Zhang et al. 2009; Di Nocera, Camilli, Terenzi 2007; Dennis and Harris 1998; van Westrenen 1996; Moroney et al. 1992; Hill et al. 1992; Battiste and Bortolussi 1988; Skipper, Rieger, and Wierwille, 1986; Casali and Wierwille 1984; Casali and Wierwille 1983). However, due to the complexity and uniqueness of many man-machine systems (Carayon et al. 2015; Dekker 2012; Carayon et al. 2006; Verwey and Veltman 1996) and human information processing (Wickens 2008), it was necessary to conduct a study which specifically

investigated their association in a modern FTD, using a realistic flying mission and operational fighter pilots.

As shown in Table 2, all MWL measures were able to differentiate the high-performance category from the medium- and low-performance categories. In addition, NASA-TLX and MCH were able to differentiate the medium- and low-performance categories. IBI was able to differentiate the rest condition from all performance categories, but was not able to differentiate the medium-performance category from the low-performance category. It is possible that some participants may have found the low-performance category trial so difficult, that they discontinued investing effort on the task. As a result, the expenditure of mental resources no longer increased with the increased task demand and the performance began to decrease (Hockey 1997; Wickens 1991; Mulder 1986; Williges and Wierwille 1979; Norman and Bobrow 1975; Kahneman 1973). To examine this possibility in more detail, the values of NASA-TLX subscales were investigated. As shown in Table 4, all NASA-TLX subscale values indicated that MWL (and its sub-dimensions) in the medium-performance category were significantly higher than in the high-performance category. However, based on the NASA-TLX subscale values there was no significant difference in effort between the medium- and low-performance categories. The lack of increased effort may partly explain the IBI values in the low-performance category. More importantly, this finding highlights the complexity of the human information processing and the added value the use of multiple MWL measures and mixed methods can provide (Carayon et al. 2015).

Even when NASA-TLX and MCH scales indicated that the low-performance category had the highest MWL, it remains debatable if there truly was a significant MWL difference between the medium-performance and low-performance categories; it is possible that the participants have interpreted poor performance as high workload

(Casner & Gore 2010). The strong, positive correlation between NASA-TLX and MCH suggests that the rating scales provide similar results. This may be a useful finding as the time required to fill NASA-TLX and MCH rating scales are different. As a result, MCH may be more appropriate option when MWL is measured in a time critical environment. In the light of the results, the main finding was that the subjective and physiological MWL measures were equally sensitive to most task demand manipulations in a modern fighter aviation domain. The possibility to use either subjective or objective MWL measures, or sometimes both, should help instructors and evaluators in customising and fine-tuning pilot training and system design.

During system design and TTP development, it is often necessary to evaluate and compare MWL both in a FTD and in real flight. Due to the limitations of the physiological MWL measures and the nature of human physiological functioning, it may be sometimes difficult or impossible to utilise physiological MWL measures. However, it is encouraging that subjective MWL measures seem to have a potential to fill this capability gap – at least when MWL is evaluated in a FTD and in a non-punitive context. The risk of potential pilot biases will still constitute a major limitation if the subjective MWL measures are extended beyond the test and evaluation settings (Gopher and Donchin 1986).

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Aasman, J., Mulder, G., and Mulder, L. J. 1987. "Operator Effort and the Measurement of Heart-Rate Variability." *Human Factors* 29 (2): 161-170.doi: 10.1177/001872088702900204.

Akyeampong, J., Udoka, S., Caruso, G., and Bordegoni, M. 2014. "Evaluation of Hydraulic Excavator Human–Machine Interface Concepts Using NASA-TLX." *International Journal of Industrial Ergonomics* 44 (3), 374-382. doi: 10.1016/j.ergon.2013.12.002.

Annett, J. 2002. "Subjective Rating Scales: Science or Art?" *Ergonomics* 45 (14), 966-987. doi: 10.1080/00140130210166951.

Battiste, V., and Bortolussi, M. 1988. "Transport Pilot Workload: A Comparison of Two Subjective Techniques." *Proceedings of the Human Factors Society Annual Meeting* 32 (2), 150-154. doi: 10.1177/154193128803200232.

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

Camm, A., Malik, M., Bigger, J., Breithardt, G., Cerutti, S., Cohen, R., Coumel, P., Fallen, E., Kennedy, H., and Kleiger, R. 1996. "Heart Rate Variability: Standards of Measurement, Physiological Interpretation and Clinical Use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology." *Circulation* 93 (5), 1043–1065. doi: 10.1161/01.CIR.93.5.1043.

Carayon, P., Hundt, A., Alvarado, C., Springman, S., and Ayoub, P. 2006. "Patient Safety in Outpatient Surgery: The Viewpoint of the Healthcare Providers." *Ergonomics* 49 (5-6), 470-485. doi: 10.1080/00140130600568717.

Carayon, P., Kianfar, S., Li, Y., Xie, A., Alyousef, B., and Wooldridge, A. 2015. "A Systematic Review of Mixed Methods Research on Human Factors and Ergonomics in Health Care." *Applied Ergonomics* 51, 291-321. doi: 10.1016/j.apergo.2015.06.001.

Casali, J., and Wierwille, W. 1983. "A Comparison of Rating Scale, Secondary-Task, Physiological, and Primary-Task Workload Estimation Techniques in a Simulated Flight Task Emphasizing Communications Load." *Human Factors* 25 (6), 623-641. doi: 10.1177/001872088302500602.

Casali, J., and 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. doi: 10.1080/00140138408963584.

Casner, S., and Gore, B. 2010. *Measuring and Evaluating Workload: A primer.* NASA Technical Memorandum (2010-216395), Ames Research Center, San Jose, CA, USA.

ChuDuc, H., NguyenPhan, K., and NguyenViet, D. 2013. "A Review of Heart Rate Variability and its Applications." *APCBEE Procedia* 7, 80-85. doi: 10.1016/j.apcbee.2013.08.016.

Dahlstrom, N., and Nahlinder, S. 2009. Mental Workload in Aircraft and Simulator During Basic Civil Aviation Training. *The International Journal of Aviation Psychology* 19 (4), 309-325. doi: 10.1080/10508410903187547.

Dekker, S. 2012. "Complexity, Signal Detection, and the Application of Ergonomics: Reflections on a Healthcare Case Study." *Applied Ergonomics* 43 (3), 468-472. doi: 10.1016/j.apergo.2011.07.003.

Dennis, K., and Harris, D. 1998. "Computer-Based Simulation as an Adjunct to Ab

Initio Flight Training." *The International Journal of Aviation Psychology* 8 (3), 261-276. doi: 10.1207/s15327108ijap0803_6.

DiNocera, F., Camilli, M., and Terenzi, M. 2007. "A Random Glance at the Flight Deck: Pilots' Scanning Strategies and the Real-Time Assessment of Mental Workload." *Journal of Cognitive Engineering and Decision Making* 1 (3), 271-285. doi: 10.1518/155534307X255627.

DiDomenico, A., and Nussbaum, M. 2008. "Interactive Effects of Physical and Mental Workload on Subjective Workload Assessment." *International Journal of Industrial Ergonomics* 38 (11), 977-983. doi: 10.1016/j.ergon.2008.01.012.

Fallahi, M., Motamedzade, M., Heidarimoghadam, R., Soltanian, A., and Miyake, S. 2016. "Effects of Mental Workload on Physiological and Subjective Responses During Traffic Density Monitoring: a Field Study." *Applied Ergonomics* 52, 95-103. doi: 10.1016/j.apergo.2015.07.009.

Gabriel, G., Ramallo, M., and Cervantes, E. 2016. "Workload Perception in Drone Flight Training Simulators." *Computers in Human Behavior*, 64, 449-454. doi: 10.1016/j.chb.2016.07.040.

Gopher, D., and Donchin, E. 1986. "Workload – an Examination of the Concept." In *Handbook of Perception and Human Performance. Cognitive Processes and Performance Vol 2,* edited by Boff, K., Kaufman, L., and Thomas, 41, 1-49. New York, NY:Wiley-Interscience.

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. doi: 10.1177/154193120605000909.

Hart, S., and Staveland, L. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." *Advances in Psychology* 52, 139-183. doi: 10.1016/S0166-4115(08)62386-9.

Hill, S., Iavecchia, H., Byers, J., Bittner, A., Zaklade, A., and Christ, R. 1992. "Comparison of Four Subjective Workload Rating Scales." *Human Factors* 34 (4), 429-439. doi: 10.1177/001872089203400405.

Hockey, R. 1997. "Compensatory Control in the Regulation of Human Performance Under Stress and High Workload: A Cognitive-Energetical Framework." *Biological Psychology* 45 (1), 73-93. doi: 10.1016/S0301-0511(96)05223-4.

Johanssen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A., and Wickens, C. 1979. "Final Report of Experimental Psychology Group." In *Mental Workload Vol. 8 of the NATO Conference Series*, 101-114. New York, NY: Springer.

Jorna, P. 1993. "Heart Rate and Workload Variations in Actual and Simulated Flight." *Ergonomics* 36 (9), 1043-1054. doi: 10.1080/00140139308967976.

Kahneman, D. 1973. *Attention and Effort*. Englewood Cliffs, NJ: Prentice Hall.

Mansikka, H., Simola, P., Virtanen, K., Harris, D., and Oksama, L. 2016. "Fighter Pilots' Heart Rate, Heart Rate Variation and Performance During Instrument Approaches." Ergonomics 59 (10), 1344-1352. doi: 10.1080/00140139.2015.1136699.

Mansikka, H., Virtanen, K., Harris, D., and Simola, P. 2016. "Fighter Pilots' Heart Rate, Heart Rate Variation and Performance During an Instrument Flight Rules Proficiency Test." *Applied Ergonomics* 56, 213-219. doi: 10.1016/j.apergo.2016.04.006.

Michell, J. 1997. "Quantitative Science and the Definition of Measurement in Psychology." *British Journal of Psychology* 88 (3), 355-383. doi: 10.1111/j.2044-8295.1997.tb02641.x.

Moroney, W., Biers, D., Eggemeier, F., and Mitchell, J. 1992. "A Comparison of Two

Scoring Procedures with the NASA Task Load Index in a Simulated Flight Task." *Proceedings of the IEEE 1992 national,* 734-740. doi: 10.1109/NAECON.1992.220513.

Mulder, G. 1986. "The Concept and Measurement of Mental Effort" In *Energetics and human information processing,* edited by Hockey, G. Dordrecht: Springer.

Newell, G., and Mansfield, N. 2008. "Evaluation of Reaction Time Performance and Subjective Workload During Whole-Body Vibration Exposure While Seated in Upright and Twisted Postures With and Without Armrests." *International Journal of Industrial Ergonomics* 38 (5), 499-508. doi: 10.1016/j.ergon.2007.08.018.

Norman, D., and Bobrow, D. 1975. "On Data-Limited and Resource-Limited Processes." *Cognitive Psychology* 7 (1), 44-64. doi: 10.1016/0010-0285(75)90004-3.

O'Donnell, R., Eggemeier, F., and Thomas, F. 1986. "Workload Assessment Methodology." In *Handbook of Perception and Human Performance. Cognitive Processes and Performance Vol 2,* edited by Boff, K., Kaufman, L., and Thomas, 42, 1-49. New York, NY:Wiley-Interscience.

Opmeer, C. 1973. "The Information Content of Successive RR-Interval Times in the ECG. Preliminary Results Using Factor Analysis and Frequency Analysis." *Ergonomics* 16 (1), 105-112. doi: 10.1080/00140137308924486.

Prichard, J., Bizo, L., and Stratford, R. 2011. "Evaluating the Effects of Team-Skills Training on Subjective Workload." *Learning and Instruction* 21 (3), 429-440. doi: 10.1016/j.learninstruc.2010.06.003.

Roscoe, A. 1975. "Heart Rate Monitoring of Pilots During Steep-Gradient Approaches." *Aviation, Space, and Environmental Medicine* 46 (11), 1410-1413.

Roscoe, A. 1993. "Heart Rate as a Psychophysiological Measure for In-Flight Workload Assessment." *Ergonomics* 36 (9), 1055-1062. doi:10.1080/00140139308967977.

Sato, N., Kamada, T., Miyake, S., Akatsu, J., Kumashiro, M., and Kume, Y. 1999. "Subjective Mental Workload in Type A Women." *International Journal of Industrial Ergonomics* 24 (3), 331-336. doi: 10.1016/S0169-8141(98)00060-2.

Skipper, J., Rieger, C., and Wierwille, W. 1986. "Evaluation of Decision-Tree Rating Scales for Mental Workload Estimation." *Ergonomics* 29 (4), 585-599. doi: 10.1080/00140138608968293.

Svensson, E., Angelborg-Thanderez, M., Sjöberg, L., and Olsson, S. 1997. "Information Complexity-Mental Workload and Performance in Combat Aircraft." *Ergonomics* 40 (3), 362-380. doi: 10.1080/001401397188206.

Svensson, E., Angelborg-Thanderz, M., and Wilson, G. 1999. *Models of Pilot Performance for Systems and Mission Evaluation - Psychological and Psychophysiological Aspect*s. *Interim Report*. Wright-Patterson Air Force Base, Dayton, OH: Human Effectiveness Directorate,.

Svensson, E. A. I., and Wilson, G. F. (2002). Psychological and psychophysiological models of pilot performance for systems development and mission evaluation. The International Journal of Aviation Psychology 12(1), 95-110.

van Westrenen, F. 1996. "A Framework for a Decision Model of River-Pilots Based on Their Workload." *International Journal of Industrial Ergonomics* 17 (5), 411-418. doi: 10.1016/0169-8141(94)00118-9.

Veltman, J., and Gaillard, A. 1998. "Physiological Workload Reactions to Increasing Levels of Task Difficulty." Ergonomics 41 (5), 656-669. doi: 10.1080/001401398186829.

Verwey, W., and Veltman, H. 1996. "Detecting Short Periods of Elevated Workload: A

Comparison of Nine Workload Assessment Techniques." J*ournal of Experimental Psychology: Applied*, 2 (3), 270-285. doi: 10.1037/1076-898X.2.3.270.

Vicente, K., Thornton, D., and Moray, N. 1987. "Spectral Analysis of Sinus Arrhythmia: A Measure of Mental Effort." *Human Factors* 29 (2), 171-182. doi: 10.1177/001872088702900205.

Wang, L., He, X., and Chen, Y. 2016. "Quantitative Relationship Model Between Workload and Time Pressure Under Different Flight Operation Tasks." *International Journal of Industrial Ergonomics* 54 93-102. doi: 10.1016/j.ergon.2016.05.008.

Wickens, C. 1991. "Processing Resources and Attention." In *Multiple-Task Performance,* edited by Damos, D., 3-34. London: Taylor and Francis.

Wickens, C. 2008. "Multiple Resources and Mental Workload." Human Factors 50 (3), 449-455. doi: 10.1518/001872008X288394.

Wierwille, W., and Eggemeier, F. 1993. "Recommendations for Mental Workload Measurement in a Test and Evaluation Environment." *Human Factors* 35 (2), 263-281. doi: 10.1177/001872089303500205.

Wierwille, W., and Casali, J. 1983. "A Validated Rating Scale for Global Mental Workload Measurement Applications." *Proceedings of the Human Factors society Annual Meeting* 27 (2), 129-133. doi: 10.1177/154193128302700203.

Wierwille, W., Rahimi, M., and Casali, J. 1985. "Evaluation of 16 Measures of Mental Workload Using a Simulated Flight Task Emphasizing Mediational Activity." *Human Factors* 27 (5), 489-502. doi: 10.1177/001872088502700501.

Williges, R., and Wierwille, W.1979. "Behavioral Measures of Aircrew Mental Workload." *Human Factors* 21 (5), 549-574. doi: 10.1177/001872087902100503.

Wilson, G. 1993. "Air-to-Ground Training Missions: A Psychophysiological Workload Analysis." *Ergonomics* 36 (9), 1071-1087. doi: 10.1080/00140139308967979.

Ylönen, H., Lyytinen, H., Leppäluoto, J., and Kuronen, P. 1997. "Heart Rate Responses to Real and Simulated BA Hawk MK 51 Flight." *Aviation, Space, and Environmental Medicine* 68 (7), 601-605.

Zhang, Y., Li, Z., Wu, B., and Wu, S. 2009. "A Spaceflight Operation Complexity Measure and its Experimental Validation." *International Journal of Industrial Ergonomics* 39 (5), 756-765. doi: 10.1016/j.ergon.2009.03.003.