Applied Ergonomics, Volume 74, January 2019, pp. 221-232
DOI: 10.1016/j.apergo.2018.08.028

Measuring mental workload using physiological measures

R L Charles, J Nixon
Cranfield University
R L Charles, Cranfield University, Martell House, Cranfield, Beds, MK43 0TR

Abstract

Technological advances have led to physiological measurement being increasingly used to measure, and predict operator states. Mental workload (MWL) in particular has been characterised using a variety of physiological sensor data. This state of the art review contributes a synthesis of the literature. We present a systematic review of 58 peer reviewed journal articles which present original data using primarily peripheral nervous system (PNS) measures to include electrocardiographic, blood pressure, respiratory, ocular and dermal sensors. In addition, electroencephalographic measures have been included if they are presented with a PNS measure. The literature reviewed covers a wide range of applied and experimental studies across various domains, with aviation being highly represented in the sample of applied literature reviewed. We present a summary of the six measures and provide a high level evidence base including how to deploy each measure, and characteristics that can affect or preclude a measure from use in a study. Measures can be used to discriminate differences in MWL caused by task type, task load, and in some cases task difficulty. In addition there is varying ranges of sensitivity to sudden or gradual changes in taskload across the six measures. We conclude that there is no single measure that clearly discriminates mental workload but there is a growing empirical basis with which to inform both science and practice.

Keywords
Mental Workload; Physiological Measures

Abbreviations

ATC	Air Traffic Control
BP	Blood Pressure
BR	Breath Rate
CNS	Central Nervous System
DBP	Diastolic Blood Pressure
ECG	Electrocardiogram
EDA	Electrodermal Activity
EDR	Electrodermal Reaction
EEG	Electroencephalogram
EOG	Electro-Oculography
ERP	Event related Brain Potentials
GSR	Galvanic skin response
HF	High frequency component of Heart Rate Variability
HR	Heart Rate
HRV	Heart Rate Variability
HTD	High Traffic Density
IBI	Interbeat interval
IWS	Integrated Workload Scale
LF	Low frequency component of Heart Rate Variability
MATB	Multi Attribute Task Battery
MF	Mid frequency component of Heart Rate Variability
MWL	Mental Workload
NN	Normal to Normal Interval
PNS	Peripheral Nervous System
PSD	Power Spectral Density
RR	R-R interval
RSME	Rating Scale of Mental Effort
SBP	Systolic Blood Pressure
SCL	Skin conductance level
SCR	Skin conductance response
SDANN	standard deviation of the NN interval
SDNN	standard deviation of the averages of NN intervals
SPL	skin potential level
SWAT	Subjective Workload Assessment Technique
TBV	Tissue Blood Volume
ULF	Ultra low frequency component of Heart Rate Variability
VACP	Visual Auditory Cognitive and Psychomotor demands

1. Introduction

Mental workload (MWL) remains an important variable with which to understand user performance (Young et al., 2014). In this article we review the evidence base for measurement of MWL using physiological measures. This review is partly in response to the array of new sensor technologies available. This field is evolving quickly, and equipment is being developed constantly that makes physiological measurement easier. Cheaper, smaller technologies allow the collection and analysis of a variety of data associated with user physiology. This data can be collected discretely and in many cases without interference with the primary task. We suggest that understanding the links between user physiology and their experience of workload can generate exciting avenues for adapting and supporting complex cognitive work in response to real-time information about user response to a task.

This idea has been progressed in the emerging field of neuroergonomics: the users themselves can deliver the cues required to optimise and select the information delivered (Parasuraman and Wilson 2008). Sensors which are able to monitor physiological variables from which workload (for example) can be characterised. This data can then be used to assess the type and extent of supporting information delivered to a user and the level at which automation, if any, is deployed (Bailey et al., 2006). Additional contextual inputs together with environmental cues could constrain or change the presentation of information depending on the user state and environmental context. Lower mental workload and higher situation awareness in response to such interventions has been evidenced in the aviation domain (Haarmann et al., 2009). In the road transport domain, evaluations of assistive technologies to establish driver distraction and to optimise presentation of information have also shown promise in terms of increases in performance and safety (Coughlin et al., 2011; Sathyanarayana et al., 2011).

Recent research shows that our ability to represent elements of the environment together with the state of the user is developing. We cannot satisfy ourselves that the single user-machine unit of analysis can characterise these rapid system-wide dynamic transformations of information across our system-of-systems. Bringing high quality contextual and user data together can inform the mode of interaction with the information sources available to the user. In stark contrast to a static, context-bound interaction, the user becomes an integral part of the way a system collaborates and assists. In this article we explore the body of scientific evidence that can assist scientists and practitioners alike to select physiological measures to assess MWL with regard to their evidence base and associated limitations.

For a concept which is intuitively appealing, a plurality of understanding about the definitions, measurement and implications of MWL exist (Young et al., 2014). This is not necessarily an issue except where different definitions are empirically compared. We suspect that the definition of MWL in research is sometimes so closely associated with its method of measurement that explicit definition is not considered and in many cases this is understandable. An important distinction that we make in this article is between taskload and workload. Taskload can be defined as the work, for example the number of tasks, performed by a user. MWL encompasses the subjective experience of a given taskload. Factors such as time constraints, environment or experience can differentiate MWL between users for the same taskload. It is possible to achieve a sense of the MWL by examination of taskload. At first glance it makes intuitive sense that the more a user must do, the higher their MWL. The higher the taskload, the higher the MWL.

However, MWL is mediated by many factors, taskload being just one. A repetitive simple task may not be cognitively challenging, but if temporal pressure is added MWL may increase affecting performance (Young et al., 2014). Conversely, a complex task may at first be perceived as challenging, and MWL experienced may be high, but through practice and experience the MWL experienced may decrease even though the taskload has not changed

In this article we treat MWL as a subjective experience in response to a taskload, which can be modified by a variety of performance shaping factors. We will return to this issue since the diversity of understanding in the literature can make comparison of scientific studies challenging.

The last systematic review of multiple physiological measures of MWL was conducted by Kramer (1990), and Jorna extensively reviewed heart rate as an index for workload (Jorna 1992). Roscoe (1992) published a review focussing specifically on pilot workload. Lean and Shan (2012) present a review focussing on Electrocardiogam (ECG) and related measures, and Electroencephalogram (EEG). More recently, Young et al., (2014) present a concise summary of physiological measures associated with MWL measurement. In this review we build on this work presenting major physiological measures used in the measurement and prediction of MWL reported in the peer reviewed literature. We systematically explore the evidence base for each measure and consider the limitations of the method itself and the empirical evidence base.

1.2 Inclusion Criteria

The following databases were searched; Pubmed, Web of Science and Google Scholar using the following terms (Physiol* AND cognitive AND workload / Physiol* AND mental AND workload), searching for peer reviewed journal articles only without any date restrictions. The results were evaluated by looking at the title and abstract, which yielded 160 papers. After looking at these in more detail and examining the keywords and references, it became clear that the initial search criteria were not sufficient to capture all relevant literature and needed refinement. After examining several literature reviews on the topic, it was identified that the physiological literature was comprised of measures of the Central Nervous System (CNS), which comprises of the brain and spinal cord or the Peripheral Nervous System (PNS) which comprises all other measures of activity such as the heart, skin and eyes. There were six main measures identified that were associated with mental workload. These are:

- Heart measures (heart rate (HR), heart rate variability (HRV) derived from ECG)
- Respiratory measures (breath rate)
- Ocular measures (blink rate, pupil size)
- Skin measures
- Blood pressure (BP)
- Brain measures (EEG, ERPs)

The literature relating to measures of the CNS is extensive, detailed and beyond the bounds of this paper. As a result, we decided to focus on the peripheral nervous system (PNS) rather than the central nervous system (CNS). EEG has been reported only when present with another method, but has not been covered extensively in this review. The revised search terms and related results are reported in table 1.

Table 1 - Summary of literature search terms

Workload + cognitive +	Heart	HRV	ECG	Respirat*	Breath*	Eye	EOG	skin	BP
Pubmed	2393	128	664	1330	865	1500	96	1168	1886
Google Scholar All	9250	451	1360	4230	3110	7930	299	3250	6530
Web of science	108	16	10	35	8	90	2	23	22
Workload + mental +	Heart	HRV	ECG	Respirat*	Breath*	Eye	EOG	skin	BP
Pubmed	4198	150	1221	2396	1485	2092	88	1766	2828
Google Scholar All	14500	719	2470	6260	4760	6210	318	4630	11100
Web of science	300	45	33	74	11	127	8	35	80

The following journals were also hand searched: Biological Psychology, Ergonomics, Applied Ergonomics, Human Factors, Aviation, Space and Environmental Medicine, International Journal of Psychophysiology, European Journal of Applied Physiology, Journal of Experimental Psychology.

Following further refinement of selection, 400 articles were then screened in detail. Only journals with an ISSN were selected. Articles were selected for inclusion in the review that that presented original research on at least one physiological measure in relation to MWL. Any studies using extensive physical activity or pharmaceutical interventions as independent variables were excluded.

Following this down selection, fifty eight articles were reviewed in detail. These articles capture a variety of sectors, measures and techniques. Measures associated with the heart, respiration, the eye, the skin and the brain were represented in the literature (table 2) Ninety three percent of studies included one or more measures associated with the heart. Sixty-six percent of the studies reviewed used a combination of physiological measures combined with subjective measures (Table 3). This alludes to the triangulation of measures used to understand user. Many safety-critical domains are represented which are frequent consumers of human factors work due to the high consequences of reduced performance. The experimental, domain-free and simulated fixed-wing operations are more represented in the sample of literature reviewed (Table 4).

Table 2 – Summary of physiological measures

Heart	Respiration	Ocular	Brain	Blood Pressure	Dermal	Other	Reference	Number of articles
54 Studies	19 Studies	28 Studies	19 Studies	10 Studies	7 Studies	3 Studies		
✓							(Braby et al., 1993; Delaney and Brodie 2000; Durantin et al., 2014; Hart and Hauser 1987; Lahtinen et al., 2007; Lee and Liu 2003; Lehrer et al., 2010; Luque-Casado et al., 2016; Mansikka et al., 2016a; Mansikka et al., 2016b; Miyake 2001; Myrtek et al., 1994; Nickel and Nachreiner 2003; Sauer et al., 2013; Schellekens et al., 2000; Tattersall and Hockey 1995; Tripathi et al., 2003)	17
✓	✓						(Backs 1994; Wu et al., 2011)	2
✓	✓			✓			(Bernardi et al., 2000)	1
✓				✓		✓	(Boutcher and Boutcher 2006)	1
		✓					(Holland and Tarlow, 1972; Recarte and Nunes, 2003; Reiner and Gelfeld, 2014)	3
✓	✓	✓	✓				(Brookings et al., 1996; Fairclough et al., 2005; Fournier et al., 1999; Sirevaag et al., 1993; Wilson and Russell 2003; Wilson 1993)	6
✓				✓			(Finsen et al., 2001; Hjortskov et al., 2004)	2
✓			✓				(Dussault et al., 2004; Dussault et al., 2005; Hsu et al., 2015)	3
✓	✓		✓				(Zhang et al., 2010)	1
✓		✓		✓			(Causse et al., 2010; Hwang et al., 2008)	2
✓		✓					(Gao et al., 2013; De Rivecourt et al., 2008; Svensson and Wilson 2009)	3
✓		✓	✓				(Fallahi et al., 2016; Hankins and Wilson 1998; Hoepf et al., 2015; Matthews et al., 2015; Ryu and Myung 2005; Wanyan et al., 2014)	6
✓		✓	✓		✓		(Wilson 2002)	1
✓	✓				✓	✓	(Miyake et al., 2009)	1
✓	✓	✓		✓			(Veltman and Gaillard 1996; Veltman and Gaillard 1998; Veltman 2002)	3
✓	✓	✓	✓		✓		(Fairclough and Venables 2006; Hogervorst et al., 2014)	2
✓	✓	✓		✓	✓	✓	(Vogt et al., 2006)	1
✓	✓	✓					(Wang et al., 2016)	1
✓	✓				✓		(Mehler et al., 2009)	1
					✓		(Collet et al., 2014)	1
							Total:	58

Table 3 – summary of studies using subjecxtive MWL measures

Measure	Reference	Number
Bedford (A. H. Roscoe and Ellis 1990)	(Braby et al., 1993; Svensson and Wilson 2009)	2
Generic WL measure	(Boutcher and Boutcher, 2006; Finsen et al., 2001; Hart and Hauser, 1987; Nickel and Nachreiner, 2003; Recarte and Nunes, 2003; Wilson, 1993, 2002)	7
NASA-TLX (Hart and Staveland 1988)	(Backs, 1994; Brookings et al., 1996; Durantin et al., 2014; Fairclough et al., 2005; Fallahi et al., 2016; Fournier et al., 1999; Gao et al., 2013; Hankins and Wilson, 1998; Hoepf et al., 2015; Hsu et al., 2015; Hwang et al., 2008; Lee and Liu, 2003; Lehrer et al., 2010; Luque-Casado et al., 2016; Matthews et al., 2015; Miyake, 2001; Miyake et al., 2009; Ryu and Myung, 2005; Sauer et al., 2013; Sirevaag et al., 1993; Svensson and Wilson, 2009; Wanyan et al., 2014)	22
RSME (Zijlstra 1993)	(Hogervorst et al., 2014; Hsu et al., 2015; De Rivecourt et al., 2008; Veltman and Gaillard 1996; Veltman and Gaillard 1998; Veltman 2002)	6
SWAT (Reid and Nygren 1988)	(Tattersall and Hockey 1995)	1
	Total:	38

Table 4 – summary of studies by experimental domain

Domain	Reference	Number
Experimental. No specific domain.	(Backs, 1994; Bernardi et al., 2000; Boutcher and Boutcher, 2006; Causse et al., 2010; Delaney and Brodie, 2000; Finsen et al., 2001; Hogervorst et al., 2014; Holland and Tarlow, 1972; Luque-Casado et al., 2016; Miyake et al., 2009; Nickel and Nachreiner, 2003; Reiner and Gelfeld, 2014; Ryu and Myung, 2005; Schellekens et al., 2000; Tripathi et al., 2003; Zhang et al., 2010)	16
Simulated fixed wing operations	(Braby et al., 1993; Dussault et al., 2005; Fairclough et al., 2005; Fairclough and Venables 2006; Fournier et al., 1999; Hsu et al., 2015; Lahtinen et al., 2007; Lee and Liu 2003; Lehrer et al., 2010; Mansikka et al., 2016a; Mansikka et al., 2016b; Miyake 2001; Prinzel et al., 2000; De Rivecourt et al., 2008; Svensson and Wilson 2009; Tattersall and Hockey 1995; Veltman and Gaillard 1996; Veltman and Gaillard 1998; Wang et al., 2016; Wanyan et al., 2014)	19
Fixed-wing operations	(Dussault et al., 2004; Hankins and Wilson 1998; Hart and Hauser 1987; Wilson 2002)	4
Remote operation of vehicles	(Durantin et al., 2014; Hoepf et al., 2015; Matthews et al., 2015; Wu et al., 2011)	4
Simulated and actual fixed-wing operations	(Veltman 2002; Wilson 1993)	2
Simulated Rotary-wing operations	(Sirevaag et al., 1993; Wilson and Russell 2003)	2

Air traffic management	(Brookings et al., 1996; Vogt et al., 2006)	2
Nuclear (energy)	(Gao et al., 2013; Hwang et al., 2008)	2
Space flight	(Sauer et al., 2013)	1
Office	(Hjortskov et al., 2004)	1
Traffic control (road)	(Fallahi et al., 2016)	1
Train Driving	(Myrtek et al., 1994)	1
Driving (automotive)	(Collet et al., 2014; Mehler et al., 2009; Recarte and Nunes 2003)	3
	Total:	58

2. Physiological Measures

2.1 Electrocardiac activity

Cardiac activity measured using ECG techniques was the most commonly used physiological measure of MWL during the search of the literature (52 reported in this paper). ECG techniques measure the electrical activity of the heart using a number of sensors. For clinical purposes up to twelve sensors can be used. Two sensors can be sufficient for consumer applications. The electrical signal from the heart is shown in Figure 1.

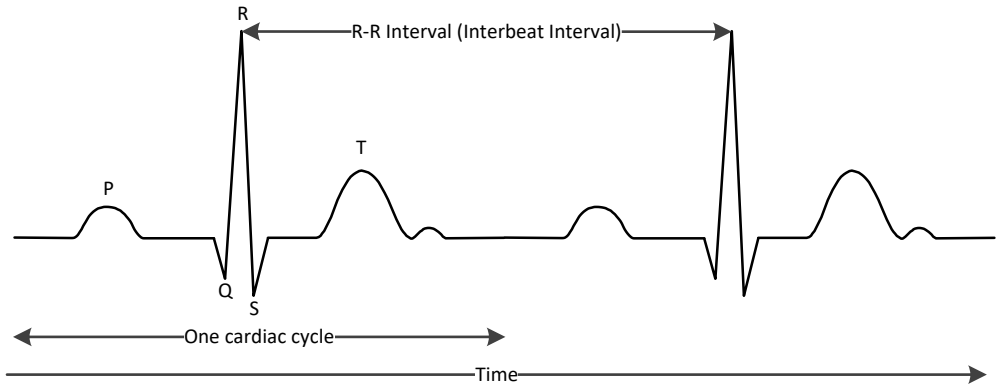

Figure 1 – The cardiac cycle

This repeating, electrical signal represents the polarisation and depolarisation of the heart required to pump blood around the body. The elements of a single cardiac cycle are termed P-Q-R-S-T. Different measures associated with this wave can be used to characterise cardiac activity giving insight into MWL. Cardiac activity can be analysed in the time or frequency domain. In either of these domains, it is Heart Rate Variability (HRV), simply the beat to beat variations, which is being characterised. Time domain measures are the simplest to perform, and Heart Rate (HR) is a popular

measurement due to its accessibility and the ability to derive several measures from the ECG. HR is typically measured as the number of beats in a period of time, most often per minute. Another measure in the time domain is the R-R interval calculated by measuring the intervals between all QRS complexes at R. This interval is known by many terms including the R-R interval, the QRS cluster interval and the interbeat-interval (IBI), all measured in milliseconds (ms). A well-used measure is the Normal to Normal (NN) interval, that also calculates the intervals at R. However, the NN interval is specifically referring to the intervals classified as normal, with abnormal beats removed. From this, the mean NN interval and other measures can be calculated, such as the standard deviation of the NN interval (SDNN) indicating variability of the NN interval. A summary of time domain measures is shown in Table 5.

Table 5 – Summary of time domain Heart Rate measures

Variable	Units	Description
NN	ms	NN interval. Also called the R-R interval or the interbeat interval (IBI). Measures the time between QRS peaks.
SDNN	ms	The standard deviation of all NN intervals.
SDANN	ms	The standard deviation of the averages of NN intervals in all 5 min segments of the entire recording.
RMSSD	ms	The square root of the mean of the sum of the squares of the differences between adjacent NN intervals.

HR increases with increasing task demands (De Rivecourt et al., 2008) and has been seen to increase during multi task conditions (Fournier et al., 1999) or when additional memory load is introduced (Finsen et al., 2001). NN intervals were also seen to decrease during a high demand multi attribute task when compared to a low demand task (Fairclough et al., 2005). Similarly, Sosnowski and colleagues (2004) saw a larger increase in HR during tasks requiring problem solving compared tasks requiring logical completion of series of elements. HR also differentiates between rest and task periods in a simulated flight task (Lahtinen et al., 2007; De Rivecourt et al., 2008), with this finding being replicated for actual flight (Dussault et al., 2004). Veltman (2002) found this change in HR to be larger during real flight when compared to simulated flight. Differences in HR have been observed during different phases of actual flight, with highest HR during take off (Hankins and Wilson 1998;

Hart and Hauser 1987; Wilson 2002) and simulated flight during ILS approach and landing (Lahtinen et al., 2007). However no significant differences in mean HR were observed between different phases of a simulated flight in Lee and Lui's (2003) study or reflected in the findings of Dussault et al., (2005). There were changes observed in these studies, but they were for more extreme changes in task load. Similarly, HR was significantly affected by task load, but only between the highest and lowest taskload conditions (Splawn and Miller 2013). Veltman and Gaillard (1998) observed a systematic decrease in NN intervals as a task became more difficult during a simulated flight task. Other cardiovascular measures differentiated between tasks but not task difficulty (Veltman and Gaillard 1998).

An increase in HR may be associated with increased visual attention, specifically with the addition of the introduction of planning tasks concurrent with flight. However, Causse et al., (2010) found HR was elevated during a logical reasoning task demanding high levels of verbal working memory compared to a dynamic reasoning task which involves planning and high visual attention, a finding also replicated by others (Boutcher and Boutcher 2006; Zhang et al., 2010). However, comparing laboratory and applied studies is difficult, as these tasks are different in nature and applied studies often use experienced participants. Ylonen et al., (1997) found higher HR in less experienced pilots in flight simulator tasks. Backs et al., (2000) observed a higher HR during a high taskload simulated air traffic control scenario when compared to a low taskload scenario but participants with no experience in air traffic control were employed.

Frequency domain methods involve defining the electocardiographic wave as different spectral components through Power Spectral Density (PSD) analysis. This analysis gives insight into the variability of the heart rate. Frequency filters are used to define the magnitude of different elements of the wave. Three spectral components commonly used. These are low frequency (LF) (0.02 - 0.06Hz), mid frequency (MF)(0.07-0.14Hz) and high frequency (HF) (0.15-0.5Hz), measured in ms^{-2}. Additionally, the LF / HF ratio may also be used. A very low frequency (VLF) component (generally

less than 0.04Hz (Henelius et al., 2009)) is sometimes used for recordings over 5 minutes (Malik et al., 1996), and when considering longer term recordings, an ultra low frequency (ULF) (less than 0.003Hz) component may also be added (Force 1996). The variance of HRV increases in line with the length of recording. For this reason, short and long term analyses of spectral components should always be distinguished. Additionally, there is variability in different studies with regards to the frequencies used. These differences make direct comparison of studies more challenging.

Veltman and Gaillard (1998) state that the MF band (0.07 – 0.14 Hz) is the most sensitive to changes in MWL, and a reduction in the power within this band reflects an increase in MWL. During a monitoring task, trainee flight engineers were observed while detecting, diagnosing and correcting faults during a simulated flight (Tattersall and Hockey 1995). They found HRV in the mid frequency band to be suppressed during the problem solving elements of the flight. Elevated HR was detected during take-off and landing phases and HRV was affected when conditions shifted from low to high traffic density during a traffic monitoring task (Fallahi et al., 2016). Increased HR was observed during a Stroop test and a decrease in the HF component was observed (Delaney and Brodie 2000), with a significant decrease in the LF band observed during high task difficulty (Delaney and Brodie 2000). This effect in the low frequency band was also observed by Splawn and Miller (2013) and Lehrer et al., (2010) at high task loads.

The HF band (0.15 – 0.50Hz) and the MF bands are affected by breathing (Veltman and Gaillard 1996), and this has been cited as a possible explanation for the finding by Gao et al., (2013) that HRV increased during the high complexity task. Deep, slow breathing during the tasks can increase HRV in the high band (Veltman and Gaillard 1998). Veltman (2002) also observed fluctuations in HRV when plotted over time that they attributed to respiratory activity, and show increases in MF variation when respiratory frequency decreases and amplitude increases. During a mental arithmetic task, NN variability increased and breathing rate decreased when speech was present (Bernardi et al., 2000), an effect that was reversed when speech was absent. In order to try and minimise this effect,

Miyake et al., (2001) instructed participants to synchronise their breathing with an audible tone during the cognitive tasks.

Backs et al., (2000) observed significant suppression in the HF band in a medium taskload condition when compared to baseline. However, this study used participants with no experience in air traffic control, so it is understandable that the results differed from Brookings et al., (1996) who found HR did not show significant differences to difficulty changes or traffic manipulation in an ATC task. However differences in in the MF band approached significance as a result of the difficulty manipulation. Dussault et al., (2005) also observed HR was lower for experts rather than novices when carrying out the same task. However, significant differences were also observed in HR and HRV between different training methods (Wu et al., 2011) which again highlights the difficulties of comparing studies that use different experience levels.

The type and length of task must be considered when comparing HRV findings. Gao et al., (2013) observed an increase in HRV during a high complexity task. This may be due to fatigue in a longer task, where HRV is seen to decrease initially then gradually increase (O'Hanlon 1972). Studies in which HRV decreased in high task load / high complexity conditions have been across shorter timescales, with the mid HRV component showing a significant increase with task duration (Fairclough et al., 2005). In addition to length of task, time of day was seen to have an effect on HRV during computer work (Schellekens et al., 2000).

Nickel et al., (2003) concluded that the MF component used to characterise HRV lacked sufficient sensitivity and diagnosticity to assess mental workload. They stated that it is only suitable for distinguishing between levels of work and levels of rest, or differences in task demand need to be high in order to be reflected in HRV (Mulder et al., 2000; Veltman and Gaillard 1998). This supports work by Braby et al., (1993) who report significant changes in HRV between an underload and a load condition during a low fidelity flight task, but not between different levels of load, and Veltman (2002) who observed that HR and HRV did not show any differences during different phases of real

and simulated flight. NN intervals and HRV were highest during rest and lowest during the parts of the task rated as more effortful by participants in a simulated flying task (Veltman and Gaillard 1998). HRV did not show significant differences to difficulty changes or traffic manipulation in an Air Traffic Control (ATC) task however HRV in the 0.15-0.4Hz band approached significance as a result of the difficulty manipulation (Brookings et al., 1996). A considerable number of studies have reported HRV in the MF band decrease when work is compared to baseline (Fallahi et al., 2016). HRV in the MF band was found to be correlated with difficulty of a tracking task (Ryu and Myung 2005) but was not found to be sensitive when a secondary mental arithmetic task was added. This does not support the findings of Fournier et al., (1999) who found that HRV in the high and medium bands was significantly lower in the multi task condition compared to the single task conditions in a multi attribute task.

The MF band has shown to be the most sensitive to task difficulty overall. However the MF band distinguishes between task demands at low to intermediate levels, but not at high levels (De Rivecourt et al., 2008). HRV in the MF band was not found to reflect differences in performance or perceived task difficulty (Nickel and Nachreiner 2003). HRV in the mid and high band also differed between segments (Veltman and Gaillard 1996). During actual flight, HRV in the medium and high bands was highly negatively correlated to HR (Hankins and Wilson 1998) but there was Lack of HRV sensitivity to task difficulty (Hankins and Wilson, 1998). Wilson (2002) concluded that HR was more sensitive than HRV for actual flight, and Veltman (2002) found that HRV did not differ between real and simulated flight during a flight study.

The key considerations of use for cardiovascular measures are domain and length of measurement. When using time domain measures, all of the recordings must be of the same length. Additionally, the cardiology task force (Force 1996) recommend that the recording length is 10 times the sampling rate. The sampling rate is important, and a range of 250-500MHz is recommended (Force 1996). There is undoubtable evidence to suggest that cardiac measures can reflect changes in levels of

work, but they are also sensitive to experience levels and training, the type of task observed and the time of day. Much of the literature has focussed on shorter term measurement of cardiac activity, which may indicate that differences are more prominent for more sudden, extreme changes in taskload. However, there is evidence to suggest that more subtle changes are reflected over time, but since shorter term differences are more significant, this could reflect publication bias.

2.2 Respiration

Respiratory measures include rate, airflow , volume, or respiratory gas analysis. For the measurement of MWL respiratory rate is the most useful of the respiratory measures (Roscoe 1992). Helpfully, respiratory rate is easy to measure through electrophysiological methods but tension measurements can also be used to gauge breath rate by placing a strap around the chest and monitoring increase and decrease in strap tension. Nineteen articles reported in this paper used respiration rate as a measure, usually derived from electrophysiological or tension methods. Airflow or gas analysis measures require a mask to be placed over the nose and mouth. While more acceptable in laboratory task conditions, these methods become less useful in applied settings and could affect primary task performance. Only one study presented here used respiratory gas analysis (Backs 1994). One exception may be military pilots who often wear oxygen masks during the course of their work, so measuring air flow and respiratory gases in this instance does not interfere with the primary task. Another consideration with respiratory measures is when speech is required. Speech production can interrupt and modify respiratory patterns leading to changes in respiratory rate unrelated to MWL (Bernardi et al., 2000; Roscoe 1992; Sirevaag et al., 1993).

Respiration rate was found to be higher as the difficulty increased during simulated Air Traffic Control (Brookings et al., 1996), a finding also observed by Backs et al., (2000). They found that Respiration rate was significantly higher during the three ATC scenarios (taskload controlled by manipulating traffic volume and traffic density) than the baseline condition and differed significantly across workload conditions. In addition, Brookings et al., (1996) observed a decrease in blink rate as

the task became more difficult and respiration rate increased. The increase in respiration rate may have been a direct result of the increased metabolic demands required to perform the task. The continuous processing that a task requires was the focus of a study by Backs et al., (1994) that used a memory task as the stimulus. They found that metabolic rate increased as the difficulty of the task increased, but also found that the metabolic rate was higher in poorer performers. Brookings (1996) did not observe any correlated changes in heart rate and respiration rate during an ATC task but respiration rate was higher as the task difficulty increased (Brookings etr al. 1996). In addition to performance, training type and length has also been shown to have an effect on respiration rate aligned to MWL (Wu et al., 2011).

During simulated aviation tasks significant increases in respiration were also reported (Fairclough and Venables 2006) with higher respiration rate observed during tasks rated subjectively as more effortful than other parts of a simulated flight task (Veltman and Gaillard 1998). Aviation tasks, by nature involve multiple demands, and these findings have been replicated in other studies involving multiple tasks such as mental arithmetic (Zhang et al., 2010) or additional mental effort with memory load and temporal demand (Backs 1994). Respiration rate increased significantly during a multi task condition compared to a single task (Fournier et al., 1999), however this was over a shorter period of time. Respiration rate was found to increase from baseline for the first 32 minutes only of a multi-attribute test, but dissipated during the final 32 minutes. Respiration rate also significantly increased during high demand compared to low demand but this effect was also only seen in the first 32 minutes (Fairclough et al., 2005)

While respiration rate has been seen to increase and respiration volume decrease as stress and workload increase, this measure is highly dependent on physical activity (Grassmann et al., 2016). As such, tasks which require exertion are not suited to this type of analysis. Additionally, tasks involving large amounts of speech should be treated with caution, as the increase in breathing rate could be a result of the increased physical output. Similar to cardiovascular measures, the length of the

recording should be taken into account when comparing studies, as the notable differences appear to level off after time. However, respiratory measures do appear to be more sensitive to gradual changes in MWL.

2.3 Skin measures

Only seven of the 58 studies considered in this paper used any sort of skin measures in relation to workload measurement or prediction. Electrodermal activity (EDA) techniques were used to assess MWL. The basis of the measurement of electrodermal activity is the change in electrical activity in the eccrine sweat glands controlled by the sympathetic nervous system. For this reason EDA is extremely sensitive to temperature, humidity, age, sex, time of day and season making comparison between studies difficult (Kramer 1990). Tissue blood volume (TBV) is also a less commonly used measure that records blood flow below the surface of the skin. TBV patterns have been found to be negatively correlated with task difficulty in a computer based task (Miyake et al., 2009). TBV is highly affected by the thermal aspects of the environment (Miyake et al., 2009). EDA can be broken down further into different components. Skin conductance level (SCL) relates to the slower characteristics of the signal. Skin conductance response (SCR) refers to faster changing elements of the signal.

During a driving task, skin conductance increased with the addition of a secondary stimulus (Mehler et al., 2009). Interestingly, skin conductance did not change when increasing the difficulty of the secondary stimulus which may indicate a lack of sensitivity in this measure when considering higher taskloads. However, electrodermal Activity (EDA) has been shown to be sensitive to sudden stimulus (Collet et al., 2014). Collet and colleagues used the duration of the electrodermal reaction (EDR) (how long EDA is detected for) as a measure. The duration of the response was found to increase as the stress increased, particularly during emergency braking.

Wilson (2002) found EDA and HR to be strongly correlated during a real flight task but not task during simulated flight. However, this study used all males, did not control for time of day and used

a small sample size (less than 10). During a computer based task, EDA correlated with task difficulty during a multi attribute task and showed better test-retest reliability than other physiological measures (Miyake et al., 2009). During a similar multi attribute task, skin conductance level increased significantly from baseline decreasing over time (Fairclough and Venables 2006). This could indicate that EDA is sensitive to sudden, but not gradual changes in MWL and has a measurement ceiling.

2.4 Blood Pressure

Blood pressure (BP) is a measure of the pressure exerted on the walls of the blood vessels by blood circulating around the body. BP is not a widely used metric for workload measurement being employed by only ten studies reviewed. BP is commonly expressed as the pressure exerted on contraction of the heart muscle (the systole) and the pressure exerted on relaxation of the heart muscle (the diastole). Pressures are measured in millimetres of mercury (mm Hg). Optimum BP is 120mm Hg systolic, 80mm Hg diastolic, abbreviated to 120/80mm Hg. Changes in BP are caused by change in sympathetic activity and is related to the mid band of HRV (Veltman and Gaillard 1996). An increase in BP has been associated with increased task load and has been shown to differentiate between periods of work and rest (Veltman and Gaillard 1996; Veltman and Gaillard, 1998) and during a simulated flight task with experienced pilots diastolic BP and BP variability was shown to differ significantly between the flight segments (Veltman and Gaillard 1996). BP was lowest during rest and highest during the task rated subjectively as more effortful (Veltman and Gaillard 1996). However, when a nuclear monitoring task became increasingly complex, BP was not reported to increase (Hwang et al., 2008). Under controlled experimental conditions, the addition of secondary tasks involving memory load to a computer task have been shown to increase BP significantly (Finsen et al., 2001). BP was elevated during a logical reasoning task demanding high levels of verbal working memory compared to a dynamic reasoning task which involves planning and high visual attention (Causse et al., 2010)

When compared with a spontaneous breathing condition, reading silently and aloud saw an increase in BP. This increase was significant during a mental arithmetic task, both when silent and aloud (Bernardi et al., 2000). Mean Arterial Pressure was observed to be higher during a traditional stroop task (verbal) when compared to a black and white, and a non verbal Stroop task. (Boutcher and Boutcher 2006). This is consistent with the MF of HRV being affected by breathing and in turn talking, which relates to BP.

Blood pressure measurements can be restrictive, especially in applied situations due to the measurement technique. BP is also heavily be influenced by state: physical activity, stress, sleep, digestion and time of day as well as the presence of speech.

2.5 Ocular Measures

Ocular measures have been used in almost half of the papers reviewed here (28) and there are a range of techniques available. The use of ocular measures has increased in recent years and this may be due to the increased ease of measurement and accessibility of apparatus. Measures include blink rate, blink duration, blink latency and pupil size.

Pupil diameter can vary from 0.2 – 0.8mm and is controlled by a group of muscles that contract and expand. The main function of this ability is to allow vision in a variety of conditions, increasing the diameter of the pupil in darker conditions and also to enable the eye to change focus (Kramer 1990). Pupil diameter has been studied in relation to MWL in both laboratory and applied studies (Kramer 1990). Mean pupil diameter change was higher during a dynamic reasoning task which involves planning and high visual attention compared to a logical reasoning task demanding high levels of verbal working memory (Causse et al., 2010) and was found to be sensitive to errors made by the participant. Pupil diameter has also been shown to reflect heart rate variations (Murata and Iwase 2000) and correlated highly with error rate during a nuclear power plant simulation (Gao et al., 2013) which may reflect an increase in MWL. The introduction of verbal outputs was seen to lead

to significant differences in pupil diameter during a real driving task (Recarte and Nunes 2003). However, care should be taken with pupil diameter measures as a decrease in pupil diameter could be the result of a decrease in ambient illumination (De Rivecourt et al., 2008).

A longer blink interval (decreased blink rate) has been observed when continued monitoring is required, for instance during a continuous tracking task in the visual modality (Ryu and Myung 2005; Stern 1980). The type of task can influence the type of changes, for example, small differences between dwell time and fixation duration during a simulated flight task may have been due to the characteristics of an instrument flight task, and the fact that the pilots scan the instruments infront of them (De Rivecourt et al., 2008). Veltman (2002) observed a large increase in blink frequency during actual flight when compared to simulated flight. This could have been for a number of reasons, including different visual stimuli in the environment or different light intensity. However blink duration decreased and amplitude increased for both the real and simulated flight.

Blink frequency and duration has been shown to decrease when participants are exposed to high visual workload and so may be used as a measure of MWL when the task of under examination is visual (Veltman and Gaillard 1996). Increased visual demand has been shown to yield lower blink rates such as during an ATC task (Brookings et al., 1996), a simulated flight task (Veltman and Gaillard 1996) actual flight (Wilson 2002) and simulated helicopter flight (Sirevaag et al., 1993). Blink rate has been shown to decrease when more visual stimuli are present and the visual demands of the task increase (Veltman and Gaillard 1996). Blink rate was shown to decrease when information was presented in the visual rather than auditory modality during a simulated helicopter flight task (Sirevaag et al., 1993) and Wilson (2002) observed decreased blink rates during visually demanding segments of actual flight. Dwell time and fixation duration was seen to decrease with increasing task demand in a simulated flight task (De Rivecourt et al., 2008). In addition, following periods of higher cognitive load, a burst of blinks may be observed (Gao et al., 2013). This may because the blinks are delayed until all the decisions relating to the external stimuli had been made (Bauer et al., 1985).

Similarly, blink rate was higher preceding incorrect responses than correct responses (Holland and Tarlow 1972). However, in order to discover this a task analysis is required to map the findings to the activity (De Rivecourt et al., 2008).

During a simulated nuclear control task, blink duration and frequency decreased during the high complexity task compared to the low complexity task (Hwang et al., 2008). This was also observed by Fairclough and Venables (2006) during a Multi Attribute Task Battery (MATB) task when compared to the baseline measures. However, these observations have mainly been short term changes and may reduce over time: Fairclough et al., (2005) report reduced blink frequency during episodes of high demand, but only for the first half of a 64 minute task.

Although ocular measures can be a good indicator of MWL, light, air quality and air conditioning or drugs can all have significant effects across all measures reported. Ocular measures can provide a good indication of when visual stimulus is dominating MWL, but again, these measures become less sensitive over time.

2.6 Brain Activity

Brain activity can be measured in the time, or frequency domain and has been a popular choice of measurement when considering workload. Nineteen studies used brain activity in this review. As stated previously, any studies using brain activity measures alone have been excluded from this review.

Event related Brain Potentials (ERPs) Are time based measures based on the occurrence of an event, whereas EEGs are decomposed in the frequency domain (see Kramer, 1990 for a description of EEG and ERPs). ERPs consist of Negative (N) or positive (P) polarity components. An N100 component would indicate a negative component occurring a minimum of 100 milliseconds after a stimulus. The P300 component is well cited in studies that use a primary task plus a secondary task to elicit the amplitude decrease (Hohnsbein et al., 1995). EEG activity is usually decompressed into frequency

bands between 1 and 40 Hz. Delta (up to 2 Hz), Theta (4 – 7Hz), Alpha (8-13Hz), and Beta (14 – 25Hz).

The P300 component has been cited as a good measure of mental workload and has been shown to decrease in amplitude when a primary task difficulty has increased. It has been cited in various domains including air traffic control (Brookings et al., 1996; Wilson and Russell 2003), flight (Hankins and Wilson, 1998; Wilson, 2002), simulated flight (Veltman and Gaillard 1996), and desk based memory (Henelius et al., 2009) mental arithmetic tasks (Henelius et al., 2009; Zhang et al., 2010), or varied tasks such as the MATB (Wilson and Russell 2003)

During a multi-task test, the p300 amplitude was found to decrease with an increasing number of simultaneous tasks (Henelius et al., 2009). Larger P300 amplitudes were also observed during low task load segments of a simulated helicopter flight task and were not affected by the type of information provision; auditory vs visual. (Sirevaag et al., 1993)

Generally, changes in task demand have been shown to lead to a change in EEG frequencies. During an ATC task, manipulation in the traffic volume and density resulted in changes in EEG frequencies (Brookings et al., 1996). Specifically, the alpha band has been found to be sensitive to memory demands (Klimesch 1997) which aligns with the findings of Ryu et al., (2005). Ryu et al., used a mental arithmetic task, requiring the participant to remember more digits in the hard condition, where alpha suppression decreased in power. Decreased alpha power was also observed during a high workload multi task test (Fournier et al., 1999). This could have been due to the increased motor activity required to control a mouse with the non-prominent hand in this task. Theta power at all sites showed significant increases as the task difficulty increased. Changes in alpha and beta band were also observed, with a decrease in alpha band activity with increased cognitive demand during simulated ATC. (Brookings et al., 1996) and various studies have found correlations between EEG increase and task difficulty with EEG correlating with the subjective reports of workload (Berka et al., 2007).

A decrease in Alpha power was also observed by Wilson (2002) during the take off and landing phases of actual flight. Activity in the alpha band was found to be sensitive to changes in WL during multiple tasks (Fournier et al., 1999) however, they did not find Alpha or theta ERPs to be sensitive to workload. Conversely, activity in the theta band, was found to negatively correlate with breath rate during a mental arithmetic task (Zhang et al., 2010). Alpha power decreased during flight when compared to control segments, but showed few significant differences between flight segments. Theta activity was shown to increase from the beginning to the end of the flight (1.5 hours) (Hankins and Wilson 1998). The type of task may account for this discrepancy, and an EEG index derived from beta/alpha plus theta was found to be able to moderate a participant's level of engagement during a multi attribute task (Prinzel et al., 2000). Power in the theta and beta bands was significantly different during the tasks of a multi attribute test when compared to baseline levels and theta activity was seen to increase during periods of high demand compared to low demand (Fairclough et al., 2005). Alpha levels were suppressed for the initial 32 mins of the task, but dissipated during the remaining 32 mins (Fairclough et al., 2005). Alpha power measures were shown to be sensitive to the differing demands of a multi task condition but were not able to distinguish between the three task load conditions (Fournier et al., 1999).

Brain activity, in particular ERPs, demonstrate a promising reflection of MWL, specifically sudden changes in levels of MWL. However, measurement and analysis of the data remains complex, hence the limited inclusion in this review.

3. Discussion

This review demonstrates that physiological measures have the power to quantify and predict MWL across a variety of domains and task types. There is no universal solution to measuring mental workload using physiological measures and no one stand out method that we could recommend following this review. However, there is an increasing body of high-quality literature that can

evidence the use of a measure and how best to deploy the measure in a study. Table 6 provides a high level summary of the evidence base available at the time of writing. Each of the findings in the table is supported by evidence in the review. The measures that the findings are supported by are indicated by ticks in the table. Overall, elements of MWL can be predicted or differentiated accross all measures considered in this review. Certain measures are more sensitive to task demands and others are more sensitive to task complexity. We have also included characteristics reported in the literature that can affect or otherwise preclude a measure from use in a study.

Table 6 - Evidence base across different measures

Finding	HRV - Time domain measures	HRV - Frequency measures				Respiration	Skin measures	BP	Pupil diameter	Blink rate	Brain activity
		HRV - VLF	HRV - LF	HRV - Mid band	HRV - high frequency						
Measure is sensitive to changes in MWL from increasing task demand	✓			✓		✓	✓	✓		✓	✓
Measures is sensitive to changes in MWL task complexity										✓	✓
Measure differentiates MWL between task type	✓			✓			✓	✓	✓	✓	
Measure differentiates MWL at extremes of taskload.	✓		✓	✓		✓					
Predictive validity of MWL is higher using tasks demanding visual attention	✓								✓	✓	
Measure is sensitive to a sudden stimulus	✓					✓	✓	✓		✓	
Measure is appropriate for shorter task duration < 5 minutes	✓		✓	✓	✓	✓	✓	✓	✓	✓	
Measure is appropriate for longer task duration > 5 minutes		✓									
Measure loses sensitivity over time						✓	✓			✓	
Measure is sensitive to errors or poor performance						✓			✓	✓	

Measure is affected by respiration				✓	✓	✓		✓			
Measure is affected by speech	✓			✓	✓	✓		✓			
Measure is affected by training and experience	✓			✓	✓	✓					
Measure is sensitive to time of day		✓	✓	✓	✓		✓	✓			
Measure is affected by ambient temperature or humidity						✓	✓	✓		✓	
Measure is affected by participant age or gender							✓	✓		✓	

Throughout the review we have found the diversity of methods and their use challenging to compare. Key issues experienced have been different perspectives and characterisations of task difficulty and task load. In addition the literature shows differences in the validity and sensitivity of measures when deployed in laboratory settings compared to real-world settings.

The way in which complexity is characterised affects the perceived workload (Park and Jung 2008; Svensson et al., 1997). Task complexity and its quantification was cited as a limitation in a study by Gao et al., (2013) where complexity was characterised numerically using the visual, auditory, cognitive and psycho-motoric (VACP) method (Aldrich et al., 1989). Other methods have included expert ratings (Lehrer et al., 2010), adding a secondary task (Durantin et al., 2014; Gao et al., 2013; Ryu and Myung 2005; Veltman and Gaillard 1996) or increasing the number of stimuli requiring action (Fournier et al., 1999; Miyake et al., 2009; Wilson and Russell 2003). The addition of rating scales which participants complete can deliver another dimension to the quantification of task complexity. However, care must be taken in applied setting when experienced operators are employed as participants since their ratings may differ significantly from those given by a less experienced participant. Additionally, the number of mistakes made by the participant can affect their subjective workload rating and can be reflected in certain physiological measures such as

respiration and eye blinks. Therefore complexity manipulation leading to poor or degraded performance should be treated with caution.

In many studies, particularly applied or simulated studies, taskload was not systematically manipulated but changed as a result of the tasks being carried out (Hankins and Wilson 1998; Lee and Liu 2003). In ATC tasks specifically, increasing traffic numbers is a popular method of increasing task load (Brookings et al., 1996). Stress levels have also been manipulated by the experimenter being 'unfriendly' to the participant during the experiment (Hjortskov et al., 2004). All of these can contribute towards a person's perception of workload as stated above. The diversity of relevant factors in ergonomics means that a body of literature may have to be very large before meaningful replication of findings and associated comparisons can be conducted. Despite the diversity, many studies use the manipulation of number of tasks to alter the level of mental workload (Henelius et al., 2009).

Comparison between laboratory and real world studies also presents challenges. Wilson (1993) found that the range of values for the same measures was much greater for the actual flight task compared to the simulated task. Changes in physiological measures cannot be easily transferred from laboratory to applied environments, and correlations have been found to be low when attempts have been made (Johnston et al., 1990; Turner and Carroll 1985). For example HR changes of up to fifty percent can be seen in applied environments, whereas they are only up to ten percent in laboratory studies (Wilson 1992).

Attempts to predict mental workload from physiological measures have had a range of success. During a computer based task, Fairclough and Venables (2006) were able to explain between one third and a half of the variance associated with the meta factor 'task engagement', with breathing rate being the most consistent; higher breathing rate meant higher task engagement. Another study combined subjective and physiological measures to give one weighted workload score. Although this score was shown to correlate with the difficulty level of certain tasks, it was concluded that

subjective workload scores may be affected greatly by the task results and the participants perceptions of good or poor performance (Miyake 2001). A correlation of 0.81 was observed between Incremental HR and NASA-TLX scores for simulated flight (Lee and Liu 2003). These two measures were deemed sensitive enough to be able to differentiate between different levels of task load, and Lehrer, (2010) found that SDNN added a 3.7% and 2.3% improvement in distinguishing high from moderate, and high from moderate and low load tasks respectively which is reflective of the literature distinguishing levels of MWL, so therefore not too surprising. Using neural networks, a range of results have been found when trying to predict MWL, but were found to be more accurate for baseline or high taskload conditions (Wilson and Russell 2003a), or load versus overload (Wilson and Russell, 2003b; Brookings et al., 1996) again reflecting the literature establing MWL using physiological measures. EEG and EOG have been found, so far to be the most promising in predicting operator state with classification accuracies of up to 90% (Hogervorst et al., 2014), and EEG has been used successfully in a closed loop design to control the tasks based upon task engagement (Pope et al., 1995).

As with most measurement strategies in human factors we have not found evidence of a silver bullet during the course of this review. No single physiological measure can provide sensitivity, diagnosticity, reliability, and ease of use. However, our review demonstrates that a strong and growing body of literature is available to the scientist and the practitioner to support the inclusion of physiological measures into human factors research. As the cost of these technologies falls or confidence in their validity rises, we hope that more individuals in the human factors community will embrace the many ways in which the measures discussed in this review can enhance the characterisation and quantification of mental workload in applied and laboratory contexts across all of the domains in which human factors currently adds so much value.

References

Aldrich, T. et al., (1989) 'The development and application of models to predict operator workload during system design', in McMillan, G. R. et al., (eds.) *Applications of Human Performance Models to System Design*. Boston, MA: Springer US, pp. 65–80.

Backs, R.W. (1994) 'Metabolic and cardiorespiratory measures of mental effort: The effects of level of difficulty in a working memory task', *International Journal of Psychophysiology*, 16(1), pp. 57–68.

Backs, R.W. et al., (2000) 'Cardiorespiratory indices of mental workload during simulated air traffic control', *Proceedings of the IEA 2000/HFES 2000 Congress*, 3, pp. 89–92.

Bailey, N.R. et al., (2006) 'Comparison of a Brain-Based Adaptive System and a Manual Adaptable System for Invoking Automation', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(4), pp. 693–709.

Bauer, L.O. et al., (1985) 'Auditory Discrimination and the Eyeblink', *Psychophysiology*, 22(6), pp. 636–641.

Berka, C. et al., (2007) 'EEG Correlates of Task Engagement and Mental\nWorkload in Vigilance, Learning, and Memory Tasks', *Aviation, Space, and Environmental Medicine*, 78(5), pp. 231–244.

Bernardi, L. et al., (2000) 'Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability', *Journal of the American College of Cardiology*, 35(6), pp. 1462–1469.

Boutcher, Y.N. and Boutcher, S.H. (2006) 'Cardiovascular response to Stroop: effect of verbal response and task difficulty.', *Biological psychology*, 73(3), pp. 235–41.

Braby, C.D. et al., (1993) 'A psychophysiological approach to the assessment of work underload', *Ergonomics*, 36(9), pp. 1035–1042.

Brookings, J.B. et al., (1996) 'Psychophysiological responses to changes in workload during simulated air traffic control', *Biological Psychology*, 42(3), pp. 361–377.

Causse, M. et al., (2010) 'Monitoring Cognitive and Emotional Processes Through Pupil and Cardiac Responses During Dynamic Versus Logical Task', *Applied psychophysiology and biofeedback*, 35(2), pp. 115–123.

Collet, C. et al., (2014) 'Measuring workload with electrodermal activity during common braking actions.', *Ergonomics*, 57(6) Taylor & Francis, pp. 886–96.

Coughlin, J.F. et al., (2011) 'Monitoring, managing, and motivating driver safety and well-being', *IEEE Pervasive Computing*, 10(3), pp. 14–21.

Delaney, J.P. and Brodie, D.A. (2000) 'Effects of short-term psychological stress on the time and frequency domains of heart-rate variability.', *Perceptual and motor skills*, 91(2) Ammons Scientific, pp. 515–24.

Durantin, G. et al., (2014) 'Using near infrared spectroscopy and heart rate variability to detect mental overload', *Behavioural Brain Research*, 259, pp. 16–23.

Dussault, C. et al., (2004) 'EEG and ECG changes during selected flight sequences', *Aviation Space and Environmental Medicine*, 75(10), pp. 889–897.

Dussault, C. et al., (2005) 'EEG and ECG changes during simulator operation reflect mental workload and vigilance', *Aviation Space and Environmental Medicine*, 76(4), pp. 344–351.

Fairclough, S.H. and Venables, L. (2006) 'Prediction of subjective states from psychophysiology: A multivariate approach', *Biological Psychology*, 71, pp. 100–110.

Fairclough, S.H. et al., (2005) 'The influence of task demand and learning on the psychophysiological response', *International Journal of Psychophysiology*, 56(2), pp. 171–184.

Fallahi, M. et al., (2016) 'Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study', *Applied Ergonomics*, 52, pp. 95–103.

Finsen, L. et al., (2001) 'Muscle activity and cardiovascular response during computer-mouse work with and without memory demands', *Ergonomics*, 44(14), pp. 1312–1329.

Force, T. (1996) 'Guidelines Heart rate variability', *European Heart Journal*, pp. 354–381.

Fournier, L.R. et al., (1999) 'Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: Manipulations of task difficulty and training', *International Journal of Psychophysiology*, 31, pp. 129–145.

Gao, Q. et al., (2013) 'Mental workload measurement for emergency operating procedures in digital nuclear power plants.', *Ergonomics*, 56(7), pp. 1070–85.

Grassmann, M. et al., (2016) 'Respiratory Changes in Response to Cognitive Load: A Systematic Review.', *Neural plasticity*.

Haarmann, A. et al., (2009) 'Combining electrodermal responses and cardiovascular measures for probing adaptive automation during simulated flight', *Applied Ergonomics*, 40(6), pp. 1026–1040.

Hankins, T.C. and Wilson, G.F. (1998) 'A Comparison of Heart Rate, Eye Activity, EEG and Subjective Measures of Pilot Mental Workload During Flight', *Aviation, Space and Environmental Medicine*, 69(4), pp. 360–367.

Hart, Sandra G. Hauser, J.R. (1987) 'Inflight Application of Three Pilot Workload Measurement Techniques', *Aviation, Space and Environmental Medicine*, 58, pp. 402–410.

Hart, S.G. and Staveland, L.E. (1988) 'Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research', *Advances in Psychology*, 52 Elsevier, pp. 139–183.

Henelius, a et al., (2009) 'Mental workload classification using heart rate metrics', *Conference proceedings : IEEE Engineering in Medicine and Biology Society*., pp. 1836–1839.

Hjortskov, N. et al., (2004) 'The effect of mental stress on heart rate variability and blood pressure during computer work', *European Journal of Applied Physiology*, 92(1-2), pp. 84–89.

Hoepf, M. et al., (2015) '*Physiological Indicators of Workload in a Remotely Piloted Aircraft Simulation'(AFRL-RH-WP-TR-2015-0092), United States Airforce*.

Hogervorst, M.A. et al., (2014) 'Combining and comparing EEG, peripheral physiology and eye-

related measures for the assessment of mental workload', *Frontiers in Neuroscience*.

Hohnsbein, J. et al., (1995) 'Effects of attention and time-pressure on P300 subcomponents and implications for mental workload research', *Biological Psychology*, 40, pp. 73–81.

Holland, M.K. and Tarlow, G. (1972) 'Blinking and Mental load', *Psychological Reports*, 31, pp. 119–127.

Hsu, B.W. et al., (2015) 'Effective Indices for monitoring mental workload while performing multiple tasks', *Perceptual and motor skills*, 121(1), pp. 94–117.

Hwang, S.L. et al., (2008) 'Predicting work performance in nuclear power plants', *Safety Science*, 46(7), pp. 1115–1124.

Johnston, D. et al., (1990) 'The relationship between cardiovascular responses in the laboratory and in the field', *Psychophysiology*, 27(1). pp. 34-44.

Jorna, P.G. (1992) 'Spectral analysis of heart rate and psychological state: a review of its validity as a workload index.', *Biological psychology*, 34, pp. 237–257.

Klimesch, W. (1997) 'EEG-alpha rhythms and memory processes', *International Journal of Psychophysiology*, 26(1-3), pp. 319–340. Available at: 10.1016/S0167-8760(97)00773-3 (Accessed: 1 June 2016).

Kramer, A.F. (1990) 'Physiological metrics of mental workload: A review of recent progress', *Multiple-task performance*, (June), pp. 279–328.

Lahtinen, T.M.M. et al., (2007) 'Heart rate and performance during combat missions in a flight simulator', *Aviation Space and Environmental Medicine*, 78(4), pp. 387–391.

Lean, Y. and Shan, F. (2012) 'Brief Review on Physiological and Biochemical Evaluatios of Human Mental Workload', *Human Factors and Ergonomics in Manufacturing*, 22(3), pp. 177–187.

Lee, Y.H. and Liu, B.S. (2003) 'Inflight workload assessment: Comparison of subjective and physiological measurements', *Aviation Space and Environmental Medicine*, 74(10), pp. 1078–1084.

Lehrer, P. et al., (2010) 'Cardiac data increase association between self-report and both expert ratings of task load and task performance in flight simulator tasks: An exploratory study', *International Journal of Psychophysiology*, 76(2), pp. 80–87.

Luque-Casado, A. et al., (2016) 'Heart rate variability and cognitive processing: The autonomic response to task demands.' *Biological psychology*, 113, pp. 83–90.

Malik, M. et al., (1996) 'Heart rate variability standards of measurement, physiological interpretation, and clinical use', *Eur Heart J*, 17(3), pp. 354–381.

Mansikka, H. et al., (2016a) 'Fighter pilots' heart rate, heart rate variation and performance during instrument approaches.', *Ergonomics*, Taylor & Francis, pp. 1–9.

Mansikka, H. et al., (2016b) 'Fighter pilots' heart rate, heart rate variation and performance during an instrument flight rules proficiency test.', *Applied ergonomics*, 56, pp. 213–219.

Matthews, G. et al., (2015) 'The Psychometrics of Mental Workload: Multiple Measures Are Sensitive

but Divergent', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(1), pp. 125–143.

Mehler, B. et al., (2009) 'Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers', *Transportation Research Record: Journal of the Transportation Research Board*, 2138 Transportation Research Board of the National Academies, pp. 6–12.

Minakuchi, E. et al., (2013) 'Evaluation of mental stress by physiological indices derived from finger plethysmography.', *Journal of physiological anthropology*, 32(1), p. 17.

Miyake, S. (2001) 'Multivariate workload evaluation combining physiological and subjective measures', *International Journal of Psychophysiology*, 40(3), pp. 233–238.

Miyake, S. et al., (2009) 'Physiological responses to workload change. A test/retest examination', *Applied Ergonomics*, 40(6), pp. 987–996.

Mulder, G. et al., (2000) 'A Psychophysiological Approach to Working Conditions', in Backs, R. W. and Boucsein, W. (eds.) *Engineering Psychophysiology: Issues and Applications*. LEA, pp. 139–159.

Murata, A. and Iwase, H. (2000) '*Evaluation Of Mental Workload By Fluctuation Analysis Of Pupil Area - Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Co*', 20(6), pp. 3094–3097.

Myrtek, M. et al., (1994) 'Physical, mental, emotional, and subjective workload components in train drivers.', *Ergonomics*, 37(7) Taylor & Francis Group, pp. 1195–203.

Nickel, P. and Nachreiner, F. (2003) 'Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload.', *Human factors*, 45(4), pp. 575–590.

O'Hanlon, J.F. (1972) '*Heart Rate Variability: A New Index of Driver Alertness/Fatigue*', SAE Technical Paper.

Parasuraman, R. and Wilson, G.F. (2008) 'Putting the brain to work: neuroergonomics past, present, and future.', *Human factors*, 50(3), pp. 468–474.

Park, J. and Jung, W. (2008) 'A study on the validity of a task complexity measure for emergency operating procedures of nuclear power plants-Comparing task complexity scores with two sets of operator response time data obtained under a simulated SGTR', *Reliability Engineering and System Safety*, 93(4), pp. 557–566.

Pope, A.T. et al., (1995) 'Biocybernetic system evaluates indices of operator engagement in automated task', *Biological Psychology*, 40, pp. 187–195.

Prinzel, L.J. et al., (2000) 'A Closed-Loop System for Examining Psychophysiological Measures for Adaptive Task Allocation A Closed-Loop System for Examining Psychophysiological Measures for Adaptive Task Allocation', *The international journal of aviation psychology*, 10(4), pp. 393–410.

Recarte, M. a and Nunes, L.M. (2003) 'Mental workload while driving: effects on visual search, discrimination, and decision making.', *Journal of experimental psychology. Applied*, 9(2), pp. 119–137.

Reid, G.B. and Nygren, T.E. (1988) 'The Subjective Workload Assessment Technique: A Scaling

Procedure for Measuring Mental Workload', *Advances in Psychology*, 52, pp. 185–218.

Reiner, M. and Gelfeld, T.M. (2014) 'Estimating mental workload through event-related fluctuations of pupil area during a task in a virtual world.', *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, 93(1), pp. 38–44.

De Rivecourt, M. et al., (2008) 'Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight.', *Ergonomics*, 51(9), pp. 1295–1319.

Roscoe, a H. (1992) 'Assessing pilot workload. Why measure heart rate, HRV and respiration?', *Biological Psychology*, 34(2-3), pp. 259–287.

Roscoe, A.H. and Ellis, G.A. (1990) '*A Subjective Rating Scale for Assessing Pilot Workload in Flight: A decade of Practical Use*' (No. RAE-TR-90019). Royal Aerospace Establishment Farnborough (united kingdom).

Ryu, K. and Myung, R. (2005) 'Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic', *International Journal of Industrial Ergonomics*, 35(11), pp. 991–1009.

Sathyanarayana, A. et al., (2011) 'Information fusion for robust "context and driver aware" active vehicle safety systems', *Information Fusion*, 12(4), pp. 293–303.

Sauer, J. et al., (2013) 'Designing automation for complex work environments under different levels of stress.', *Applied ergonomics*, 44(1), pp. 119–27.

Schellekens, J.M. et al., (2000) 'Immediate and delayed after-effects of long lasting mentally demanding work', *Biological Psychology*, 53(1), pp. 37–56.

Sirevaag, E.J. et al., (1993) 'Assessment of pilot performance and mental workload in rotary wing aircraft', *Ergonomics*, 36(9), pp. 1121–1140.

Sosnowski, T. et al., (2004) 'Program running versus problem solving: Mental task effect on tonic heart rate', *Psychophysiology*, 41(3), pp. 467–475.

Splawn, J.M. and Miller, M.E. (2013) 'Prediction of perceived workload from task performance and heart rate measures', *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting*., pp. 778–782.

Stern, J.A. (1980) '*Aspects of Visual Search Activity Related to Attentional Processes and Skill Development*' Electromagnetic Technology Corp. Paulo Alto, CA.

Svensson, E. et al., (1997) 'Information complexity--mental workload and performance in combat aircraft.', *Ergonomics*, 40(3), pp. 362–380.

Svensson, E.A.I. and Wilson, G.F. (2009) 'Psychological and Psychophysiological Models of Pilot Performance for Systems Development and Mission Evaluation', *The International Journal of Aviation Psychology*, 12(1) Lawrence Erlbaum Associates, Inc., pp. 95–110.

Tattersall, A.J. and Hockey, G.R.J. (1995) 'Level of Operator Control and Changes in Heart Rate Variability during Simulated Flight Maintenance', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(4), pp. 682–698.

Tripathi, K.K. et al., (2003) 'Attentional modulation of heart rate variability (HRV) during execution of PC based cognitive tasks', *Industrial Journal of Aerospace Medicine2*, 47(1), pp. 1–10.

Turner, J. and Carroll, D. (1985) 'The relationship between laboratory and "real world"heart rate reactivity: an exploratory study', *Journal of cardiovascular control: models, methods and data,* NATO Conference Series, III, Human Factors. Vol. 26.

Veltman, J. a. and Gaillard, a. W.K. (1998) 'Physiological workload reactions to increasing levels of task difficulty', *Ergonomics*, 41(5), pp. 656–669.

Veltman, J.A. (2002) 'A comparitive Study of Psychophysiological Reactions During Simulator and Real Flight', *The International Journal of Aviation Psychology*, 12(1), pp. 33–48.

Veltman, J.A. and Gaillard, W.K. (1996) 'Physiological indices of workload in a simulated flight task', *Biological Psychology*, 42(3), pp. 323–342.

Vogt, J. et al., (2006) 'The impact of workload on heart rate and blood pressure in en-route and tower air traffic control', *Journal of Psychophysiology*, 20(4), pp. 297–314.

Wang, Z. et al., (2016) 'Physiological Indices of Pilots' Abilities Under Varying Task Demands.', *Aerospace medicine and human performance*, 87(4) Aerospace Medical Association, pp. 375–81.

Wanyan, X. et al., (2014) 'Improving pilot mental workload evaluation with combined measures.', *Bio-medical materials and engineering*, 24(6) IOS Press, pp. 2283–90.

Wilson, G.F. (1993) 'Air-to-ground training missions: a psychophysiological workload analysis', *Ergonomics*, 36(9), pp. 1071–1087.

Wilson, G.F. (2002) 'An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures', *The international journal of aviation psychology*, 12(1), pp. 3–18.

Wilson, G.F. (1992) 'Applied use of cardiac and respiration measures: Practical considerations and precautions', *Biological Psychology*, 34(2-3), pp. 163–178.

Wilson, G.F. and Russell, C.A. (2003a) 'Real-time assessment of mental workload using psychophysiological measures and artificial neural networks.', *Human factors*, 45(4), pp. 635–643.

Wilson, G.F. and Russell, C.A. (2003b) 'Operator functional state classification using multiple psychophysiological features in an air traffic control task', *Human factors*, 45(3), pp. 381–389.

Wu, B. et al., (2011) 'Using Physiological Parameters to Evaluate Operator's Workload in Manual Controlled Rendezvous and Docking (RVD)', *Technology*, , pp. 426–435.

Ylönen, H. et al., (1997) 'Heart rate responses to real and simulated BA Hawk MK 51 flight.', *Aviation, Space and Environmental Medicine*, 68(7), pp. 601–605.

Young, M.S. et al., (2014) 'State of science: mental workload in ergonomics.', *Ergonomics*, 58(1) pp. 1–17.

Zhang, J. et al., (2010) 'Effects of mental tasks on the cardiorespiratory synchronization.', *Respiratory physiology & neurobiology*, 170(1), pp. 91–5.

Zijlstra, F. (1993) *Efficiency in work behaviour*. Technical University, Delft, The Netherlands.

2018-09-13

Measuring mental workload using physiological measures: a systematic review

Charles, Rebecca L.

Elsevier

Charles R and Nixon J. (2019) Measuring mental workload using physiological measures: a systematic review. Applied Ergonomics, Volume 74, January 2019, pp. 221-232
https://doi.org/10.1016/j.apergo.2018.08.028