Applied Ergonomics 74 (2019) 221–232

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

Review article

Measuring mental workload using physiological measures: A systematic review

Rebecca L. Charles*, Jim Nixon

Cranfield University, Martell House, Cranfield, Bedford, MK43 0TR, United Kingdom

ARTICLE INFO

Keywords:
Mental workload
Taskload
Physiological measures
Systematic review

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 systematic review contributes a synthesis of the literature summarising key findings to assist practitioners to select measures for use in evaluation of MWL. We also describe limitations of the methods to assist selection when being deployed in applied or laboratory settings.

We detail fifty-eight peer reviewed journal articles which present original data using physiological measures to include electrocardiographic, respiratory, dermal, blood pressure and ocular. Electroencephalographic measures have been included if they are presented with another measure to constrain scope. The literature reviewed covers a range of applied and experimental studies across various domains, safety-critical applications being highly represented in the sample of applied literature reviewed. We present a summary of the six measures and provide an evidence base which includes how to deploy each measure, and characteristics that can affect or preclude the use of a measure in research. Measures can be used to discriminate differences in MWL caused by task type, task load, and in some cases task difficulty. Varying ranges of sensitivity to sudden or gradual changes in taskload are also evident 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.

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 and more mobile (Nixon and Charles, 2017). Cheaper, smaller technologies allow the collection and analysis of a variety of data associated with user physiology (Guzik and Malik, 2016). These data can be collected unobtrusively 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 (for example see Christensen and Estepp, 2013). We systematically review and present evidence that can assist scientists and practitioners alike to select physiological measures to assess MWL in an evidence based way. Many of the measures have limitations that preclude their use in certain tasks or applied settings and this review will help when selecting measures for a chosen task or experiment. Finally, we summarise the key findings of the review in a table which can be used to guide experimental design or to select measures for a particular task or application.

For a concept which is intuitively appealing, a plurality of understanding about the definitions and measurement of MWL exist (Young et al., 2014). This plurality was explored by a significant workshop hosted by Neville Moray and subsequent publication of the sessions in 1979 (Moray, 1979). Moray characterises the different attributes of operator workload from a variety of different perspectives throughout the system. More recently Pickup et al. (2005) detail the difficulties in distinguishing where the influence of workload and its measurement in the system is located. Workload is not only multidimensional in nature (Xie and Salvendy (2000) but is also experienced by the operator and imposed by the task as demand. Workload can be imagined as an input and an output, being both experienced by the user subjectively, demanded of the user by the work and expended by the user to do work.

* Corresponding author.
E-mail address: r.l.charles@cranfield.ac.uk (R.L. Charles).

https://doi.org/10.1016/j.apergo.2018.08.028
Received 1 September 2016; Received in revised form 16 May 2018; Accepted 30 August 2018
Available online 13 September 2018
0003-6870/

Abbreviations

ATC	Air Traffic Control
BP	Blood Pressure
ECG	Electrocardiogram
EDA	Electrodermal Activity
EDR	Electrodermal Reaction
EEG	Electroencephalogram
EOG	Electrooculogram
ERP	Event Related Potentials
HF	High Frequency
HR	Heart Rate
HRV	Heart Rate Variability
IBI	Interbeat Interval
LF	Low Frequency
MATB	Multi-Attribute Task Battery
MF	Mid Frequency
MWL	Mental Workload
N	Negative
NASA	National Aeronautics and Space Agency
NN	Normal Normal
P	Positive
PSD	Power Spectral Density
RMSSD	Root Mean Square Standard Deviation
RSME	Rating Scale of Mental Effort
SCR	Skin Conductance Response
SDANN	Standard Deviation of the Average Normal Normal [interval]
SDNN	Standard Deviation of the Normal Normal [interval]
SWAT	Subjective Workload Assessment Technique
TBV	Tissue Blood Volume
TLX	Task Load Index
ULF	Ultra Low Frequency
VACP	Visual Auditory Cognitive Psychomotor
VLF	Very Low Frequency

These different elements of workload are individually and interactively valid depending on the questions being asked and by whom. From a psychological background, mental workload may be framed using cognitive psychology in terms of task switching or allocation of attention (Wickens, 2008). A system designer may describe workload in terms of demand placed on the user by the system or what work is required of the operator. One user may experience workload very differently to another due to individual differences (Grassmann et al., 2017). Sharples and Megaw (2015) have updated the discussion placing operator workload at the centre of a framework which includes both the physical and cognitive task demands, the operator performance and other external or internal factors. The complex interactions suggested by the framework may give rise to challenges inherent in the measurement of workload and any theoretical framework used to underpin conclusions or make predictions in this space.

Diverse perspectives as to the nature of workload and its measurement may not be issues in themselves when an experiment or task is bounded. Internal validity may be claimed. Where trouble can emerge is through the formal comparison of studies employing different definitions, measurements or constructs relating to mental workload. It is for this reason that our original ambition to conduct a formal meta-analysis of the studies was rejected. Notwithstanding the power of meta-analyses to cope with differing methodologies, the diversity of theoretical treatments and task types we have observed would have rendered any conclusions unreliable at best. 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 (Matthews et al., 2015). Researchers or practitioners may satisfy themselves with the face or content validity of a reliable measurement instrument without needing to explore theoretical underpinnings.

To locate this review amid this diversity, we distinguish 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 (Sharples and Megaw, 2015; Wickens, 2008). 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 (Colle and Reid, 1998). 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 (see Matthews et al., 2015). In this review we treat MWL as a subjective experience in response to a taskload, which can be modified by a variety of performance shaping factors.

The last review of multiple physiological measures of MWL was conducted by Kramer (1990). Jorna also 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 electrocardiogram (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 expand the range of measures considered and review the recent evidence base for measurement of mental workload using major physiological measures reported in the peer-reviewed literature across multiple domains. We systematically explore the evidence base for each measure and consider the limitations

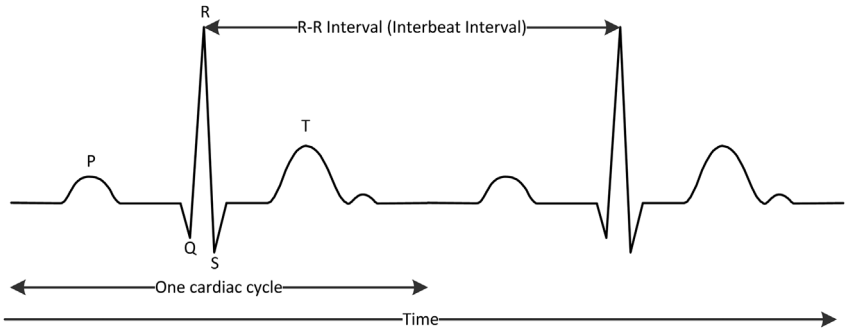

Fig. 1. The cardiac cycle.

of the method itself and the empirical evidence base. Six key measures are considered in this review: electrocardiac activity, respiration, skin-based measures, blood pressure, ocular measures and electrical brain activity. These measures are characterised both by the representation in the published, peer-reviewed literature and by their relative ease of measurement. For example, measurement of these physiological parameters does not require invasive, medical procedures or very expensive static equipment. An introduction to each measurement method considered in this review is detailed below.

1.1. Electrocardiac activity

ECG techniques measure the electrical activity of the heart using a number of sensors. The electrical signal from the heart is shown in Fig. 1. For clinical purposes up to twelve sensors can be used. Two sensors can be sufficient for consumer applications.

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), the beat to beat variations, which is being characterised. Time domain measures are the more straight forward measures to compute. A summary of time domain measures is shown in Table 1. Heart rate (HR) is a typically reported measurement and is measured as the number of beats in a period of time, most often reported 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 widely 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 which is not the case for the R-R interval. 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.

Frequency domain methods involve defining the electrocardiographic wave as different spectral components through power spectral density (PSD) analysis, typically by application of a Fourier transform. Frequency filters are used to define the magnitude of different elements of the wave over a period of time. Three spectral components are commonly used. These are low frequency (LF) (0.02–0.06 Hz), mid frequency (MF) (0.07–0.14 Hz) and high frequency (HF) (0.15–0.5 Hz). Additionally, the LF/HF ratio may also be used. A very low frequency (VLF) component, generally less than 0.04 Hz is sometimes used for recordings over 5 min (Henelius et al., 2009; Malik et al., 1996). When considering longer-term recordings, an ultra-low frequency (ULF) of less than 0.003 Hz component may also be added (AHA and ESC, 1996).

1.2. 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. Airflow or gas analysis measures require a mask to be placed over the nose and mouth.

1.3. Skin based measures

The basis of the measurement of electrodermal activity (EDA) is the change in electrical activity in the eccrine sweat glands controlled by the sympathetic nervous system. For this reason EDA is sensitive to temperature, humidity, age, sex, time of day and season, making comparison between studies particularly 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.

1.4. 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 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 120 mm Hg systolic, 80 mm Hg diastolic, abbreviated to 120/80 mm Hg.

1.5. Ocular measures

The use of ocular measures has increased in recent years and this may be due to t increased ease of measurement and accessibility of apparatus in recent years. Measures include blink rate, blink duration, blink latency and pupil size. Pupil diameter can vary from two to eight millimetres 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).

1.6. Brain measures

Brain activity can be measured in the time, or frequency domain and has been a popular choice of measurement when considering workload. In this review we only consider brain activity which is electrically sensed. Methods such as functional magnetic resonance imaging (fMRI) currently require expensive, fixed based equipment. As such their deployment in applied settings is not currently possible. Event related Brain Potentials (ERPs) are electrically sensed based on the time following the occurrence of an event. ERPs consist of Negative (N) or positive (P) polarity components. An N100 component would indicate a negative component occurring a minimum of 100 ms after a stimulus. The P300 component is cited in workload studies that use a primary task plus a secondary task to elicit the amplitude decrease (Hohnsbein et al., 1995). An electroencephalogram (EEG) is an electrically sensed signal over time that can be decomposed in the frequency domain (see Kramer, 1990 for a description of EEG and ERPs). EEG activity is often decomposed into frequency bands between 1 and 40 Hz. Delta (up to 2 Hz), Theta (4 – 7 Hz), Alpha (8–13 Hz), and Beta (14–25 Hz). Brain activity, in particular ERPs, demonstrate a promising reflection of

Table 1
Summary of time domain heart rate measures.

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

MWL, specifically abrupt changes in levels of MWL. However, measurement and analysis of the data remains complex, hence the scoped inclusion in this review.

This review is structured around the measures described above. Firstly, we detail the inclusion criteria for the articles reviewed. Within this section we present the articles included and the physiological measures used to assess MWL. We summarise the subjective measures employed by the articles included and the domain of study. Secondly, we review the evidence base of each measure and describe key studies that support the measurement of MWL across the range of task types and domains. We summarise the key findings in a table which gives an instant overview of an extant evidence base across the measures considered. Finally, we discuss key themes that have emerged from the review.

2. Inclusion criteria

The Pubmed, Web of Science and Google Scholar databases were searched. Terms include an asterisk representing a wild character. The search terms and related results are reported in Table 2. Initially, the following terms were employed: (physiol* AND cognitive AND workload/physiol* AND mental AND workload). Only peer reviewed journal articles were included without any date restrictions. The results were evaluated by examination of the title and abstract yielding 160 papers. After looking at these in more detail and examining the keywords and references, six main measures were identified associated with mental workload: heart measures (heart rate [HR], heart rate variability [HRV] derived from ECG), respiratory measures (breath rate), skin measures, blood pressure (BP), ocular measures (blink rate, pupil size) and brain measures (EEG). These terms were then included in the search. The literature relating to EEG measures is extensive, detailed and beyond the scope of this paper. These measures have been reported only when present with another method.

Journals that were heavily represented in this first set of searches were selected for further scrutiny. Using the search terms, manual searching using the journal homepage and in some cases paper copies of the journals held in the library were conducted. This process ensured that earlier work was assured as included in the searches. The following journals were treated in this way: Biological Psychology, Ergonomics, Applied Ergonomics, Human Factors, Aviation, Space and Environmental Medicine, International Journal of Psychophysiology, European Journal of Applied Physiology, and the Journal of Experimental Psychology.

Following further refinement of selection, 400 articles were then screened in detail. Articles were selected for inclusion in the review that that presented original research of 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 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 skin, the eye and the brain were represented in the literature (Table 3) 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 4). This alludes to the triangulation of measures used to understand user mental workload emphasised by Sharples and Megaw (2015). 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 5).

3. Evidence for physiological measures

3.1. 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). 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 et al. (2004) saw a larger increase in HR during tasks requiring problem solving compared to tasks requiring logical completion of a 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 precision-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 larger 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

Table 2
Summary of literature search terms.

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

Table 3
Representation of physiological measures across the literature.

Heart (54)	Respiration (19)	Skin (7)	Blood Pressure (10)	Ocular (28)	Brain (19)	Other (3)	References	Number of articles
✓							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
✓	✓			✓	✓		Brookings et al., 1996; Fairclough et al., 2005; Fournier et al., 1999; Sirevaag et al., 1993; Wilson and Russell 2003a,b; Wilson 1993	6
✓				✓	✓		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
				✓			Holland and Tarlow, 1972; Recarte and Nunes, 2003; Reiner and Gelfeld, 2014	3
✓					✓		Dussault et al., 2004; Dussault et al., 2005; Hsu et al., 2015	3
✓				✓			Gao et al., 2013; De Rivecourt et al., 2008; Svensson and Wilson 2009	3
✓	✓		✓	✓			Veltman and Gaillard 1996; Veltman and Gaillard 1998; Veltman 2002	3
✓	✓						Backs 1994; Wu et al., 2011	2
✓			✓				Finsen et al., 2001; Hjortskov et al., 2004	2
✓			✓	✓			Causse et al., 2010; Hwang et al., 2008	2
✓	✓	✓		✓	✓		Fairclough and Venables 2006; Hogervorst et al., 2014	2
✓	✓		✓				Bernardi et al., 2000	1
✓			✓			✓	Boutcher and Boutcher 2006	1
✓	✓				✓		Zhang et al., 2010	1
✓		✓		✓	✓		Wilson 2002	1
✓	✓	✓				✓	Miyake et al., 2009	1
✓	✓	✓	✓	✓		✓	Vogt et al., 2006	1
✓	✓			✓			Wang et al., 2016	1
✓	✓	✓					Mehler et al., 2009	1
		✓					Collet et al., 2014	1

Table 4
Studies which employ physiological measures *and* at least one subjective MWL measure.

Measure	Reference	Number
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
Bespoke 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
Rating Scale of Mental Effort (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
Bedford scale workload assessment (Roscoe and Ellis, 1990)	Braby et al., 1993; Svensson and Wilson 2009	2
SWAT (Reid and Nygren, 1988)	Tattersall and Hockey, 1995	1

Table 5
Domain used in articles reviewed.

Domain	Reference	Number
Simulated aviation operations (fixed wing)	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
Non-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
Aviation operations (fixed-wing)	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
Automotive	Collet et al., 2014; Mehler et al., 2009; Recarte and Nunes 2003	3
Simulated and real fixed-wing aviation operations	Veltman 2002; Wilson 1993	2
Simulated rotary-wing operations	Sirevaag et al., 1993; Wilson and Russell 2003a,b	2
Air traffic management	Brookings et al., 1996; Vogt et al., 2006	2
Nuclear power	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
Rail	Myrtek et al., 1994	1

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.

In the frequency domain, 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 (0.15–0.50 Hz) 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 (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 the MF band approached significance as a result of the manipulation of difficulty. Dussault et al. (2005) also observed HR was lower for experts compared to 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 and Nachreiner (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 0.15–0.4 Hz 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 reduced 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 (AHA and ESC, 1996) recommend that the recording length is 10 times the sampling rate. The sampling rate is important, and a range of 250–500 MHz is recommended (AHA and ESC, 1996). In the frequency domain, 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.

3.2. Respiration

Nineteen articles reported in this paper used respiration rate as a measure, usually derived from electrophysiological or tension methods. Respiration rate was found to be higher as the difficulty increased during simulated ATC (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 et al. (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 et 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 min only of a multi-attribute test, but this change dissipated during the final 32 min. Respiration rate also significantly increased during high demand compared to low demand but this effect was also only seen in the first 32 min (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. Gas analysis methods which demand that masks are worn over the mouth and nose are less useful in applied settings and could affect primary task performance. 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. Only one study presented here used respiratory gas analysis (Backs, 1994). 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).

3.3. Skin measures

Only seven of the 58 studies considered in this paper used any type of skin measures in relation to workload measurement or prediction. 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) 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 of fewer than ten participants. 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 evidence could indicate that EDA is sensitive to sudden, but not gradual changes in MWL and has a measurement ceiling.

3.4. Blood pressure

BP is not a widely used metric for workload measurement being employed by ten of the studies reviewed here. BP is heavily influenced by state: physical activity, stress, sleep, digestion and time of day as well as the presence of speech (Adams and Leverland, 1985). 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, 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 speech production, which relates to BP.

3.5. Ocular measures

Ocular measures have been used in nearly half of the papers reviewed here (28) and there are a range of techniques available. 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 change 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 in front 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 have 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 and response (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 min 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.

3.6. Brain activity

Nineteen studies used brain activity in this review. As stated previously, any studies using brain activity measures alone have been excluded from this review. The P300 component has been cited as a reliable 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, 2003a; b), flight (Hankins and Wilson, 1998; Wilson, 2002), simulated flight (Veltman and Gaillard, 1996), and desk based task (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, 2003a; b).

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 and Myung (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 breathing 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 h) (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 min of the task, but dissipated during the remaining 32 min (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).

4. Summary of physiological measures findings

In this section the evidence reviewed is summarised and provides an overview of the measures in relation to each of the findings or claims identified in the literature. Table 6 provides a high level summary of the evidence base available at the time of writing. The measures that support the findings or claims are indicated by ticks in the table. Each of the findings in the table is supported by at least one paper in the review, without any conflicting research at the time of writing. A dash (−) in a cell indicates that the finding is not evidenced in the peer-reviewed literature at the time of writing. Presenting the findings in this way provides a quick reference guide for practitioners and scientists that can be referred to when designing experiments or selecting the most appropriate measure. The table will also be of value when selecting measures for applied settings since characteristics such as time of day may not be controlled.

5. Discussion

This review details and demonstrates a growing empirical basis for the use of physiological measures in quantifying MWL across a variety of domains and task types. There is no universal solution to measuring mental workload using physiological measures and no single 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 inform how best to deploy the measure in a study.

The different perspectives and characterisations of task difficulty and task load have been a key challenge to comparing and synthesising studies in this review. In addition the literature shows differences in the validity and sensitivity of measures when deployed in laboratory settings compared to real-world settings adding another layer of complexity this task.

The complexity of tasks is particularly difficult to assess and compare between applied studies. 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 psychomotor (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;

Table 6
Summary of findings supported by the literature for each physiological measure reviewed.

Finding/claim	Time domain HRV	VLF HRV	LF HRV	MF HRV	HF HRV	Respiration	Skin measures	BP	Pupil diameter	Blink rate	Brain activity
Measure differentiates MWL between higher or lower taskload.	✓	–	✓	✓	–	✓	–	–	–	–	–
Measure differentiates MWL between task types.	✓	–	–	✓	–	–	✓	✓	✓	✓	–
Measure is sensitive to changes in MWL from increasing task demand.	✓	–	–	✓	–	✓	✓	✓	–	✓	✓
Predictive validity of MWL is higher using tasks demanding visual attention.	✓	–	–	–	–	–	–	–	✓	✓	–
Measure is sensitive to a sudden stimulus.	✓	–	–	–	–	✓	✓	✓	–	✓	–
Measure is affected by ambient temperature or humidity.	–	–	–	–	–	✓	✓	✓	–	✓	–
Measure is sensitive to time of day.	–	✓	✓	✓	✓		✓	✓	–	–	–
Measure is appropriate for shorter task duration (< 5 min).	✓	–	✓	✓	✓	✓	✓	✓	✓	✓	–
Measure is appropriate for longer task duration (> 5 min).	–	✓	–	–	–	–	–	–	–	–	–
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 affected by participant age or gender.	–	–	–	–	–	–	✓	✓		✓	–

Veltman and Gaillard, 1996) or increasing the number of stimuli requiring action (Fournier et al., 1999; Miyake et al., 2009; Wilson and Russell, 2003a; b). 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 settings 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 varied 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 participant's 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. 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 many measurement strategies in human factors we have not found evidence of a silver bullet during the course of this review. We cannot currently argue that any individual physiological parameter can provide a single true measure of MWL experienced in response to a task. That said, physiological measures do capture the experience of the user in response to task demands. This is in contrast to subjective measures of workload which can become confounded with perceptions of task demand or task performance (Pickup et al., 2005). However, this is of course on the understanding that the physiological measures themselves are capturing the experience of mental workload in a reliable way.

Our review does demonstrate 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.

Acknowledgements

This article is based on work performed in the programme: Future Sky Safety, which has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 640597. The views and opinions expressed in this paper are those of the authors and are not intended to represent the position or opinions of the Future Sky Safety consortium or any of the individual partner organisations.

References

Adams, C.E., Leverland, M.B., 1985. Environmental and behavioral factors that can affect blood pressure. 1985 Nov. Nurse Pract. Am. J. Prim Health Care 10 (11) 39-40, 49-50.
AHA, ESC, 1996. Guidelines for Heart rate variability. Eur. Heart J. 354–381.
Aldrich, T., et al., 1989. The development and application of models to predict operator workload during system design. In: McMillan, G.R. (Ed.), Applications of Human Performance Models to System Design. Springer US, Boston, MA, 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. Int. J. Psychophysiol. 16 (1), 57–68.
Backs, Richard W., Ryan, Arthur M., Wilson, Glenn F., 1994. Psychophysiological measures of workload during continuous manual performance. Hum. Factors 36 (3), 514–531.
Backs, R.W., et al., 2000. Cardiorespiratory indices of mental workload during simulated air traffic control. In: Proceedings of the IEA 2000/HFES 2000 *Congress*, vol. 3. pp. 89–92.
Bauer, L.O., et al., 1985. Auditory discrimination and the eyeblink. Psychophysiology 22 (6), 636–641.
Berka, C., et al., 2007. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviat Space Environ. Med. 78 (5), 231–244.
Bernardi, L., et al., 2000. Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability. J. Am. Coll. Cardiol. 35 (6), 1462–1469.
Boutcher, Y.N., Boutcher, S.H., 2006. Cardiovascular response to Stroop: effect of verbal response and task difficulty. Biol. Psychol. 73 (3), 235–241.
Braby, C.D., et al., 1993. A psychophysiological approach to the assessment of work underload. Ergonomics 36 (9), 1035–1042.
Brookings, J.B., Wilson, G.F., Swain, C.R., 1996. Psychophysiological responses to changes in workload during simulated air traffic control. Biol. Psychol. 42 (3), 361–377.
Causse, M., et al., 2010. Monitoring cognitive and emotional processes through pupil and cardiac responses during dynamic versus logical task. Appl. Psychophysiol. Biofeedback 35 (2), 115–123.
Christensen, James C., Estepp, Justin R., 2013. Coadaptive aiding and automation enhance operator performance. Hum. Factors 55 (5), 965–975.
Colle, H., Reid, G.B., 1998. Context effects in subjective mental workload ratings. Hum. Factors 40 (4), 591–600.
Collet, C., et al., 2014. Measuring workload with electrodermal activity during common braking actions. Ergonomics 57 (6), 886–896 Taylor & Francis.
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), 1295–1319.
Delaney, J.P., Brodie, D.A., 2000. Effects of short-term psychological stress on the time and frequency domains of heart-rate variability. Percept. Mot. Skills 91 (2), 515–524 Ammons Scientific.
Durantin, G., et al., 2014. Using near infrared spectroscopy and heart rate variability to detect mental overload. Behav. Brain Res. 259, 16–23.
Dussault, C., et al., 2004. EEG and ECG changes during selected flight sequences. Aviat Space Environ. Med. 75 (10), 889–897.
Dussault, C., et al., 2005. EEG and ECG changes during simulator operation reflect mental

workload and vigilance. Aviat Space Environ. Med. 76 (4), 344–351.
Fairclough, S.H., Venables, L., 2006. Prediction of subjective states from psychophysiology: a multivariate approach. Biol. Psychol. 71, 100–110.
Fairclough, S.H., et al., 2005. The influence of task demand and learning on the psychophysiological response. Int. J. Psychophysiol. 56 (2), 171–184.
Fallahi, M., et al., 2016. Effects of mental workload on physiological and subjective responses during traffic density monitoring: a field study. Appl. Ergon. 52, 95–103.
Finsen, L., et al., 2001. Muscle activity and cardiovascular response during computer-mouse work with and without memory demands. Ergonomics 44 (14), 1312–1329.
Fournier, L.R., et al., 1999. Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training. Int. J. Psychophysiol. 31, 129–145.
Gao, Q., et al., 2013. Mental workload measurement for emergency operating procedures in digital nuclear power plants. Ergonomics 56 (7), 1070–1085.
Grassmann, Mariel, Vlemincx, Elke, von Leupoldt, Andreas, Mittelstädt, Justin M., Van den Bergh, Omer, 2016. Respiratory changes in response to cognitive load: a systematic review. Neural Plast. 2016, 8146809 16 pages. https://doi.org/10.1155/2016/8146809.
Grassmann, M., Vlemincx, E., von Leupoldt, A., Van den Bergh, O., 2017. Individual differences in cardiorespiratory measures of mental workload: an investigation of negative affectivity and cognitive avoidant coping in pilot candidates. Appl. Ergon. 59, 274–282. http://doi.org/10.1016/j.apergo.2016.09.006.
Guzik, Przemyslaw, Malik, Marek, 2016. ECG by mobile technologies. J. Electrocardiol. 49 (6), 894–901.
Hankins, T.C., Wilson, G.F., 1998. A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. Aviat Space Environ. Med. 69 (4), 360–367.
Hart, Sandra G., Hauser, J.R., 1987. Inflight application of three pilot workload measurement techniques. Aviat Space Environ. Med. 58, 402–410.
Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183 Elsevier.
Henelius, a, et al., 2009. Mental workload classification using heart rate metrics. In: *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. Eur. J. Appl. Physiol. 92 (1–2), 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. Front. Neurosci. 8, 322.
Hohnsbein, J., et al., 1995. Effects of attention and time-pressure on P300 subcomponents and implications for mental workload research. Biol. Psychol. 40, 73–81.
Holland, M.K., Tarlow, G., 1972. Blinking and mental load. Psychol. Rep. 31, 119–127.
Hsu, B.W., et al., 2015. Effective Indices for monitoring mental workload while performing multiple tasks. Percept. Mot. Skills 121 (1), 94–117.
Hwang, S.L., et al., 2008. Predicting work performance in nuclear power plants. Saf. Sci. 46 (7), 1115–1124.
Johnston, D., et al., 1990. The relationship between cardiovascular responses in the laboratory and in the field. Psychophysiology 27 (1), 34–44.
Jorna, P.G., 1992. Spectral analysis of heart rate and psychological state: a review of its validity as a workload index. Biol. Psychol. 34, 237–257.
Klimesch, W., 1997. EEG-alpha rhythms and memory processes. Int. J. Psychophysiol. 26 (1–3), 319–340. https://doi.org/10.1016/S0167-8760(97)00773-3 , Accessed date: 1 June 2016.
Kramer, A.F., 1990. Physiological metrics of mental workload: a review of recent progress. Multiple-Task Perform. 279–328 June.
Lahtinen, T.M.M., et al., 2007. Heart rate and performance during combat missions in a flight simulator. Aviat Space Environ. Med. 78 (4), 387–391.
Lean, Y., Shan, F., 2012. Brief review on physiological and biochemical evaluatios of human mental workload. Hum. Factors Ergon. Manuf. 22 (3), 177–187.
Lee, Y.H., Liu, B.S., 2003. Inflight workload assessment: comparison of subjective and physiological measurements. Aviat Space Environ. Med. 74 (10), 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. Int. J. Psychophysiol. 76 (2), 80–87.
Luque-Casado, A., et al., 2016. Heart rate variability and cognitive processing: the autonomic response to task demands. Biol. Psychol. 113, 83–90.
Malik, M., et al., 1996. Heart rate variability standards of measurement, physiological interpretation, and clinical use. Eur. Heart J. 17 (3), 354–381.
Mansikka, H., et al., 2016a. Fighter pilots' heart rate, heart rate variation and performance during instrument approaches. Ergon., Taylor & Francis 1–9.
Mansikka, H., et al., 2016b. 'Fighter pilots' heart rate, heart rate variation and performance during an instrument flight rules proficiency test. Appl. Ergon. 56, 213–219.
Matthews, G., Reinerman-Jones, L.E., Barber, D.J., Abich, J., 2015. The psychometrics of mental workload: multiple measures are sensitive but divergent. Hum. Factors: J. Hum. Factors Ergon. Soc. 57 (1), 125–143.
Mehler, B., et al., 2009. Impact of incremental increases in cognitive workload on physiological arousal and performance in Young adult drivers. Transport. Res. Rec.: J. Transport Res. Board 6–12 2138 Transportation Research Board of the National Academies.
Miyake, S., 2001. Multivariate workload evaluation combining physiological and subjective measures. Int. J. Psychophysiol. 40 (3), 233–238.
Miyake, S., et al., 2009. Physiological responses to workload change. A test/retest examination. Appl. Ergon. 40 (6), 987–996.
Moray, N.E., 1979. Mental Workload: its Theory and Measurement. Plenum Press, New York.
Mulder, G., et al., 2000. A psychophysiological approach to working conditions. In: Backs, R.W., Boucsein, W. (Eds.), Engineering Psychophysiology: Issues and Applications. LEA, pp. 139–159.
Murata, A., Iwase, H., 2000. Evaluation of mental workload by fluctuation analysis of pupil area - engineering in medicine and biology society, 1998. Proc. 20th Ann. Int. Co 20 (6), 3094–3097.
Myrtek, M., et al., 1994. Physical, mental, emotional, and subjective workload components in train drivers. Ergonomics 37 (7), 1195–1203 Taylor & Francis Group.
Nickel, P., Nachreiner, F., 2003. Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. Hum. Factors 45 (4), 575–590.
Nixon, J., Charles, R., 2017. Understanding the human performance envelope using electrophysiological measures from wearable technology. Cognit. Technol. Work 19 (4), 655–666.
O'Hanlon, J.F., 1972. Heart Rate Variability: a New Index of Driver Alertness/Fatigue (SAE Technical Paper). .
Park, J., 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. Reliab. Eng. Syst. Saf. 93 (4), 557–566.
Pickup, L., Wilson, J.R., Sharples, S., Norris, B., Clarke, T., Young, M.S., 2005. Fundamental examination of mental workload in the rail industry. Theor. Issues Ergon. Sci. 6 (6), 463–482. http://doi.org/10.1080/14639220500078021.
Pope, A.T., et al., 1995. Biocybernetic system evaluates indices of operator engagement in automated task. Biol. Psychol. 40, 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. Int. J. Aviat. Psychol. 10 (4), 393–410.
Recarte, M. a, Nunes, L.M., 2003. Mental workload while driving: effects on visual search, discrimination, and decision making. J. Exp. Psychol. Appl. 9 (2), 119–137.
Reid, G.B., Nygren, T.E., 1988. The subjective workload assessment technique: a scaling procedure for measuring mental workload. Adv. Psychol. 52, 185–218.
Reiner, M., Gelfeld, T.M., 2014. Estimating mental workload through event-related fluctuations of pupil area during a task in a virtual world. Int. J. Psychophysiol.: Official J. Int. Organ. Psychophysiol. 93 (1), 38–44.
Roscoe, a H., 1992. Assessing pilot workload. Why measure heart rate, HRV and respiration? Biol. Psychol. 34 (2–3), 259–287.
Roscoe, A.H., 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., Myung, R., 2005. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. Int. J. Ind. Ergon. 35 (11), 991–1009.
Sauer, J., et al., 2013. Designing automation for complex work environments under different levels of stress. Appl. Ergon. 44 (1), 119–127.
Schellekens, J.M., et al., 2000. Immediate and delayed after-effects of long lasting mentally demanding work. Biol. Psychol. 53 (1), 37–56.
Sharples, S., Megaw, T., 2015. Definition and measurement of human workload. In: Wilson, J.R., Sharples, S. (Eds.), Evaluation of Human Work, fourth ed. CRC Press, London, pp. 515–548.
Sirevaag, E.J., et al., 1993. Assessment of pilot performance and mental workload in rotary wing aircraft. Ergonomics 36 (9), 1121–1140.
Sosnowski, T., et al., 2004. Program running versus problem solving: mental task effect on tonic heart rate. Psychophysiology 41 (3), 467–475.
Splawn, J.M., Miller, M.E., 2013. Prediction of perceived workload from task performance and heart rate measures. In: 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.A.I., Wilson, G.F., 2009. Psychological and psychophysiological models of pilot performance for systems development and mission evaluation. Int. J. Aviat. Psychol. 12 (1), 95–110 Lawrence Erlbaum Associates, Inc.
Svensson, E., et al., 1997. Information complexity–mental workload and performance in combat aircraft. Ergonomics 40 (3), 362–380.
Tattersall, A.J., Hockey, G.R.J., 1995. Level of operator control and changes in heart rate variability during simulated flight maintenance. Hum. Factors: J. Hum. Factors Ergon. Soc. 37 (4), 682–698.
Tripathi, K.K., et al., 2003. Attentional modulation of heart rate variability (HRV) during execution of PC based cognitive tasks. Ind. J. Aero. Med. 47 (1), 1–10.
Turner, J., Carroll, D., 1985. The relationship between laboratory and "real world"heart rate reactivity: an exploratory study. J. Cardiovasc. Contr.: Model. Meth. Data 26 NATO Conference Series, III, Human Factors.
Veltman, J.A., 2002. A comparitive study of psychophysiological reactions during simulator and real flight. Int. J. Aviat. Psychol. 12 (1), 33–48.
Veltman, J.A., Gaillard, W.K., 1996. Physiological indices of workload in a simulated flight task. Biol. Psychol. 42 (3), 323–342.
Veltman, J. a., Gaillard, a. W.K., 1998. Physiological workload reactions to increasing levels of task difficulty. Ergonomics 41 (5), 656–669.
Vogt, J., et al., 2006. The impact of workload on heart rate and blood pressure in en-route and tower air traffic control. J. Psychophysiol. 20 (4), 297–314.
Wang, Z., et al., 2016. Physiological indices of pilots' abilities under varying task demands. Aero. Med. Hum. Perform. 87 (4), 375–381 Aerospace Medical Association.
Wanyan, X., et al., 2014. Improving pilot mental workload evaluation with combined measures. Bio Med. Mater. Eng. 24 (6), 2283–2290 IOS Press.
Wickens, Christopher D., 2008. Multiple resources and mental workload. Hum. Factors 50

(3), 397–403.
Wilson, G.F., 1992. Applied use of cardiac and respiration measures: practical considerations and precautions. Biol. Psychol. 34 (2–3), 163–178.
Wilson, G.F., 1993. Air-to-ground training missions: a psychophysiological workload analysis. Ergonomics 36 (9), 1071–1087.
Wilson, G.F., 2002. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. Int. J. Aviat. Psychol. 12 (1), 3–18.
Wilson, G.F., Russell, C.A., 2003a. Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Hum. Factors 45 (4), 635–643.
Wilson, G.F., Russell, C.A., 2003b. Operator functional state classification using multiple psychophysiological features in an air traffic control task. Hum. Factors 45 (3), 381–389.
Wu, B., et al., 2011. Using Physiological Parameters to Evaluate Operator's Workload in Manual Controlled Rendezvous and Docking (RVD). pp. 426–435 Technology.
Xie, B., Salvendy, G., 2000. Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. Work. Stress 14 (1), 74–99. http://doi.org/10.1080/026783700417249.
Ylönen, H., et al., 1997. Heart rate responses to real and simulated BA Hawk MK 51 flight. Aviat Space Environ. Med. 68 (7), 601–605.
Young, M.S., et al., 2014. State of science: mental workload in ergonomics. Ergonomics 58 (1), 1–17.
Zhang, J., et al., 2010. Effects of mental tasks on the cardiorespiratory synchronization. Respir. Physiol. Neurobiol. 170 (1), 91–95.
Zijlstra, F., 1993. Efficiency in Work Behaviour. Technical University, Delft, The Netherlands.