Measuring Mental Workload With Low-Cost and Wearable Sensors: Insights Into the Accuracy, Obtrusiveness, and Research Usability of Three Instruments

Julia C. Lo, Emdzad Sehic, Prorail, and **Sebastiaan A. Meijer**, KTH Royal Institute of Technology

The affordability of wearable psychophysiological sensors has led to opportunities to measure the mental workload of operators in complex sociotechnical systems in ways that are more objective and less obtrusive. This study primarily focuses on the sensors themselves by investigating low-cost and wearable sensors in terms of their accuracy, obtrusiveness, and usability for research purposes. Two sensors were assessed on their accuracy as tools to measure mental workload through heart rate variability (HRV): the E3 from Empatica and the emWave Pro from HeartMath. The BioPatch from Zephyr Technology, which is an U.S. Food and Drug Administration–approved device, was used as a gold standard to compare the data obtained from the other 2 devices regarding their accuracy on HRV. Linear dependencies for 6 of 10 HRV parameters were found between the emWave and BioPatch data and for 1 of 10 for the E3 sensor. In terms of research usability, both the E3 and the BioPatch had difficulty acquiring either sufficiently high data recording confidence values or normal distributions. However, the BioPatch output files do not require postprocessing, which reduces costs and effort in the analysis stage. None of the sensors was perceived as obtrusive by the participants.

Keywords: cognitive processes, topics, command and control, domains, ground transportation, domains, workload, topics, analysis methods, topics

Address correspondence to Julia Lo, Department of Innovation and Development, ProRail, Moreelsepark 3, 3511 EP Utrecht, The Netherlands, Julia.Lo@prorail.nl.

Journal of Cognitive Engineering and Decision Making
2017, Volume 11, Number 4, December 2017, pp. 323–336
DOI: 10.1177/1555343417716040

INTRODUCTION

Over the last few decades, there has been a significant shift of focus from operators' physical demands to their cognitive demands: a shift that is especially pertinent in complex and safety-critical systems, such as aviation, driving, and rail (Young, Brookhuis, Wickens, & Hancock, 2014). These cognitive demands are reflected in an operator's mental workload (MWL), or the proportion of mental capacity on a task (Brookhuis & de Waard, 2010; Kahneman, 1973).

A variety of MWL measurement techniques can be divided into three general categories: performance, subjective, and physiological measurement techniques (Brookhuis, 2004; Cain, 2007; Wierwille & Eggemeier, 1993). Performance-based techniques are assessed through the capability in which an operator is able to perform system or task functions, such as speed and accuracy on primary and secondary tasks (Wierwille & Eggemeier, 1993). Second, subjective measurement techniques rely on operator judgements. Self-rating scales of MWL—such as the NASA Task Load Index by Hart and Staveland (1988), the Subjective Workload Assessment Technique by Reid and Nygren (1988), the Overall Workload Scale by Vidulich and Tsang (1987), the Modified Cooper-Harper Scale by Wierwille and Casali (1983), and the Rating Scale Mental Effort by Zijlstra (1985)—are a few of well-known examples that are used in multiple domains (Young, Brookhuis, Wickens, & Hancock, 2014). From a neuroergonomic approach, psychophysiological measurements of workload focus on the physiological responses of operators. Eye blinks through eye-tracking measurements, brain activity through electroencephalogram (EEG), and electrodermal activity

(EDA) based on sweat glands are a few examples of measurement techniques to gain insights into the MWL of operators (Brookhuis & de Waard, 2010; G. F. Wilson, 2002). Another assessment technique is that of heart rate variability (HRV; Mulder, De Waard, & Brookhuis, 2004; Parasuraman, 2011). Here the variable durations between heartbeats can have different oscillations patterns (Brookhuis, 2004).

So far, the results from multiple studies that compare measurement techniques are scattered; there is no measurement technique that has been unanimously acknowledged to be valid and reliable for MWL (Cain, 2007). In a recent study, multiple psychophysiological measurement techniques have been used to compare MWL, such as EEG, electrocardiogram (ECG), transcranial Doppler sonography, functional near infrared, and eye tracking (Matthews, Reinerman-Jones, Barber, & Julian Abich, 2015). The findings from this study indicate that certain metrics are more sensitive to dual tasking—namely, short fixation duration and a high task load index (derived from EEG). Metrics more sensitive to workload-associated change detection included, for example, a low HRV (derived from ECG) and a high theta (derived from EEG).

In the current study, MWL measurement through HRV is further explored due to the increasing availability of wearable and consumer devices and the potential advantages of these low-cost and/or less intrusive devices for human factors research.

Psychophysiological tools focusing on HRV measurements have traditionally been conducted with three-lead ECG sensors attached to the chest. The removal of bodily hair is a basic procedure to conduct the measurements, often followed by a laboratory or simulator setting in which participants are restricted in their movements by a wired connection to the recording system. Over the past few years, "wearable" sensors have become more widely available. These sensors are unobtrusive in the sense that they can be worn as an ambulatory system in which the individual is able to move around freely during his or her measurement. The emergence of wearable sensors can be explained by recent advances in microelectronics that have overcome the limitations resulting from the size of front-end electronics and the sensor itself (Patel, Park, Bonato, Chan, & Rodgers, 2012). Wearable sensors in the form of rings, wristbands, earplugs, and so on, are increasingly used by individuals to monitor their health (e.g., during workouts), whereas sensors in the form of patches are more commonly used by medical practitioners for the remote monitoring of their patients. With consumers and medical practitioners as potential clients in the heart rate–monitoring market, these devices became more affordable than ECG sensors used for research.

These innovative measurement devices have great potential for research in the railway domain, where human factors research has been steadily increasing since the railway sector, in many countries worldwide, was broken up into commercial and governmental organizations in the 1990s (J. R. Wilson & Norris, 2005). In the Netherlands, radical changes in the railways are being implemented, as long-term targets have been set to increase the infrastructural capacity (Meijer, 2012). In 2014, financial investments made by the government for capacity expansion projects amounted to only 600 million euros, which is 53% of the total funding (ProRail, 2015). Projects focusing on meeting a higher-capacity demand vary from optimizations of physical infrastructural configurations, such as creating more buffers in bottleneck areas to process optimizations. These changes also affect railway traffic operations, for which it is therefore essential to investigate the impact on the cognitive demands imposed on railway traffic operators.

A series of railway gaming simulation sessions have been conducted to test future modes of the railway system, which is expected to carry increasing amounts of freight and numbers of passengers each year (Lo, Van den Hoogen, & Meijer, 2013). Insights into the MWL of railway traffic operators are considered valuable in these sessions, with a finite pool of operators, who often return as participants. Less obtrusive instruments, such as wearable sensors, could facilitate data collection without requiring too much effort on the part of the participating operators such that they may not need be interrupted during their tasks. Also, the use of instruments

that are less obtrusive can support conditions where participants are more aware of being in a simulated environment or an investigation. Furthermore, less invasive instruments can be helpful when investigations on operators are not an established routine, as in the railway sector, in which case individuals might be more prone to feelings of hesitance and reluctance to participate. Additionally, longer MWL measurement periods during operators' shifts in real-world settings can be obtained. Current measurement techniques in the railway domain obtain estimates of MWL through self-rating or observation-based rating tools (e.g., Pickup et al., 2005). As such, the use of objective methods, such as psychophysiological measurement techniques, can provide insights into MWL development next to the existing tools.

In addition, due to their easy application and user-friendly interfaces for data extraction and visualizations, low-cost and wearable psychophysiological sensors can reduce the need for assistance from expert researchers (who are usually required when traditional heart rate measurement instruments are used) and thus make data processing in large-scale research efforts much easier.

The present study focused on establishing which low-cost and/or wearable sensors are suitable for accurate MWL measurements and, therefore, future deployment. Three psychophysiological sensors that had not yet been researched were selected from a broad range of low-cost and wearable sensors: the E3 from Empatica (Milan, Italy), the emWave Pro from HeartMath (Boulder Creek, CA), and the Bio Patch from Zephyr (Annapolis, MD). They were used to measure the MWL of train traffic controllers in a simulator through the analysis of HRV. To identify the suitability of sensors for analysis and future deployment, the focus was on identifying (1) the extent to which participants perceived the sensors as being obtrusive; (2) the sensors' usability for research purposes, in terms of usable data points and ease of post-processing data for the analysis; and (3) the accuracy of the E3 and emWave sensors versus the BioPatch ECG sensor as a U.S. Food and Drug Administration (FDA)–approved gold standard reference device.

The following section presents a description of the three psychophysiological instruments and their specifications, followed by the method, results, and discussion and conclusion.

PSYCHOPHYSIOLOGICAL INSTRUMENTS

The possibility of applying low-cost wearable sensors (e.g., Garmin Edge 800, Polar S810, and Suunto t6) to measure HRV has been explored in a number of studies, mostly by comparing the sensor to a gold standard (typically an ECG; Essner, Sjöström, Ahlgren, & Lindmark, 2013; Hlotova, Cats, & Meijer, 2014; Porto & Junqueira, 2009; Schäfer & Vagedes, 2013; Wallén, Hasson, Theorell, Canlon, & Osika, 2012; Weippert et al., 2010). Although their findings are inconclusive, these studies indicate that the results from wearable sensors are significant when physical activity, gender, and age are taken into account and care is taken in the use of absolute values. This section provides a description of the functionalities and specifications of the E3, emWave, and BioPatch devices, which were selected on the basis of placement at different locations on the body—respectively, at the wrist, ear, and chest—their low-cost availability (EmWave), their recommended application for research purposes (E3 and BioPatch), and the possibility to extract data sets from the devices. Table 1 shows the similarities and differences among the three sensors (Empatica, 2013; Zephyr, 2013).

E3

Empatica's E3 device (see Figure 1) is a wristband that has a photoplethysmograph sensor, an EDA sensor, a three-axis accelerometer, and a temperature sensor (Garbarino, Lai, Bender, Picard, & Tognetti, 2014). Photoplethysmograph is often used in wrist sensors to assess variations in the reflected or transmitted light to measure blood volume pulse, which in turn can be used to derive heart rate and HRV.

The measurements are started by pressing a button on the sensor. To access the raw data, it first has to be uploaded from the device to the Empatica website.

A unique feature of Empatica sensors is that they include an EDA sensor for skin conductance.

TABLE 1: Specifications of the E3, emWave Pro, and BioPatch Sensors

	E3	emWave Pro	BioPatch
Sampling rate, Hz	64	1	250–1,000
Max recording, hr	32	1	35
Real-time data display	iPhone app	PC software	Android app
Data extraction format	.csv	.txt,.json	.csv, .dat, .hed
Price based on list prices and quotations, U.S. dollars	1,100	299	449

Figure 1. E3 wristband from Empatica.

Figure 2. emWave Pro from HeartMath.

EDA is known to be correlated with an individual's emotional state (e.g., stress) and workload, as perspiration increases skin conductance (Pina, Donmez, & Cummings, 2008). The sensor, which is available in various sizes, is fixed to the wristband.

For the current study, only HRV analyses were conducted. To obtain the HRV data set from the E3, the data from the device need to be uploaded to the Empatica website. Empatica uses algorithms to provide online visualizations of the data and the option to download data sets with the interbeat intervals (IBIs)—that is, the time intervals between heartbeats. In the data set, a time stamp for the first recorded data point is provided with, subsequently, the number of seconds after the start of the recording and its respective IBIs. Note that the raw data are not accessible nor downloadable.

emWave Pro

The emWave Pro is offered as a training system for individuals to achieve better emotional balance. The sensor is attached to the earlobe and features only a photoplethysmograph sensor to measure heart rate (see Figure 2). As the sensor is wired to the measurement computer via a USB connector, it is not considered a wearable device.

The emWave Pro product version is required to access the raw data from the recorded sessions. Measurements and raw file extractions are accessible with the emWave software. The data set is obtained through the software program and provides the initial time stamp with the start of the recording time of the data. The data set consists of IBI data per second.

BioPatch

The BioPatch by Zephyr Technology is an FDA-approved ECG sensor that can be placed on the chest with two replaceable electrodes (see Figure 3). FDA approval indicates that the sensor complies with regulations on users' safety and is therefore permitted for use in a clinical setting. The BioPatch is rather unobtrusive in comparison with the standard three-lead ECG sensor, as no chest hairs need to be removed (depending on its placement).

The BioPatch is capable of measuring heart rate, interbeat (difference between R-wave occurrence times [RR]) interval, respiration rate, ECG,

Figure 3. BioPatch from Zephyr Technology.

activity level, and posture (Zephyr, 2013). The measurement is triggered by pressing a button on the sensor. To obtain the raw data from the recorded sessions, the BioPatch needs to be connected to a computer and accessed through a log downloader. HRV data can be analyzed through the ECG data set file or the IBI data set file. In the present study, the sampling rate of the BioPatch was set to the default 250 Hz, and the ECG data set file was used to conduct the analyses.

METHOD

Experimental Setting

A human-in-the-loop simulator that is used to train traffic controllers and test infrastructural, timetable, and process changes was deployed to conduct the experiments. This simulator, called the PRL game (version 3.5.7), incorporates a subset of the traffic management system's features. PRL is an acronym for "Proces Leiding" in Dutch, which translates to the name of the traffic management system. For this experiment, its features were extended and matched to the level of validity of the simulator for the current task at hand, in which operators had an interface design highly similar to their regular workstation. Table 2 shows the simulator design characteristics based on a gaming simulation design format (Lo & Meijer, 2013).

Scenarios

Each 20-min scenario comprised 4 phases, in which the start of a phase was triggered by an event (e.g., a call from a train driver). Each phase approximately took 5 min. These scenarios and phases were developed and selected by a subject matter expert. Scenario 1 focused on an overloaded freight train that needs to drive with a lower speed, therefore delaying subsequent trains. In Scenario 2, a malfunctioning freight train blocks multiple tracks near a railroad crossing, triggering a railroad crossing malfunction in time. In Table 3, the four phases of this scenario are summarized.

Participants

Twelve male train traffic controllers took part in the simulator session. Their average work experience was 14 years (SD = 8.5).

Materials

Perceived obtrusiveness. The degree of obtrusiveness of the sensors was qualitatively measured through questions during the debriefing. Two questions were asked to measure the degree of perceived obtrusiveness: "Did you experience the sensors as obtrusive?" "Which sensor did you find the least obtrusive?" The data were processed by checking how many participants answered the first question with yes or no and which sensor was indicated as being least obtrusive.

Usability of the sensors for research. The usability of the sensors was assessed by the extent to which they were suitable for research purposes in terms of usable data points and the ease of postprocessing data for the analysis. The postprocessing of data was necessary to identify the start and end of a selected period or phase. Reducing postprocessing saves time and reduces the costs involved in the analysis stage of a study. In terms of usable data points, the collected data were evaluated on their heart rate confidence values and normal distribution,

TABLE 2: Simulator Design Characteristics

Core aspect	Description
Purpose	Studying the impact of a new infrastructure and a new train timetable on the workload of train traffic controllers
Scenarios	Two scenarios: (1) 2014 train timetable and infrastructure and (2) 2015 train timetable and infrastructure
Simulated world	Detailed infrastructure; detailed timetable of the Nijmegen workstation; additional safety-critical features; communication possible with train drivers, the regional network controller of Arnhem, and other roles
No. of participants	One per session
Roles	Train traffic controller
Type of role	Similar to one's own role
Objectives	Execution of tasks—similar as in one's daily work
Constraints	Exclusion of a number simulator features
Load	Medium disruptions in both scenarios
Situation (external influencing factors)	Presence of three facilitators in the room
Time model	Continuous

which also indicates the quality of the obtained data.

Sensor accuracy. In line with the notion of conducting analyses without the assistance of expert researchers who use specific programming tools, HRV was analyzed in each phase with a freeware software packages called Kubios HRV 2.2. Kubios has been applied in numerous clinical studies (Fagundes et al., 2011). Based on a visual inspection of the outliers, medium artifact corrections were applied for the emWave and E3, and a very low artifact correction was applied to the BioPatch data. HRV parameters were similarly calculated for all three sensors by the time-domain method (mean difference between R-wave occurrence times; also known as mean RR) or IBI in milliseconds, standard deviation of normal-to-normal RR intervals, root mean square of successive differences in milliseconds, relative amount of successive intervals differing >50 ms in percentages (pNN50), triangular interpolation in milliseconds, and frequency-domain method (lower frequency [LF] band: 0.04–0.15 Hz and higher frequency [HF] band: 0.15–0.4 Hz, as calculated by Kubios in terms of power, normalized values, and ratio). For an elaborate description see, for example, Schäfer and Vagedes (2013) and Tarvainen, Niskanen, Lipponen, Ranta-aho, and Karjalainen (2014).

Procedure. In this within-subjects design, the participants were briefed on the purpose of the simulator study at the start of each session, and their permission was obtained to use the data from the session anonymously. All participants gave their consent and were then equipped with all three sensors. The E3 was worn on the wrist that was not used for writing or handling the computer mouse. The BioPatch was placed on the sternum on the first day. During the study, it was decided to change the location of the BioPatch for the second measurement day, to the rib left of the sternum (as seen from the rear of the participant), to increase the data-recording quality. The emWave sensor was applied to the left earlobe, due to the location of the computer. As the train traffic controllers' tasks did not require physical activities, movement activities were limited to seating positions and arm movements.

After the final scenario, the participants were debriefed about their experience with the sensors, their MWL, and their usage of the simulator. Questionnaires were handed out prior to and after the session and after each scenario.

TABLE 3: Phase Descriptions Across Scenarios 1 and 2

Phase	Trigger	Description	Expected mental workload
		Scenario 1	
1	Start of the scenario	The train traffic controller is building up his or her situation awareness and monitors the current train traffic flow.	Low
2	Regional network controller calls	Request is received to make a change in the order of trains. The operator needs to mitigate delays by manually managing the train traffic flow.	Medium
3	Train driver calls	Request is received whether the freight train has a subsequent route without a stop, as this will cause additional delay. The operator needs to wage the consequences of this request.	Medium
4	Train driver calls	Information request is received from a train driver regarding the reason for the red signal that he or she encountered.	Low
		Scenario 2	
5	Start of the scenario	The train traffic controller is building up his or her situation awareness and monitors the current train traffic flow.	Low
6	Train traffic controller performs first safety procedure	A freight train has a malfunction, blocking multiple tracks that can cause a possible rail-crossing failure. Safety procedure needs to be performed where the train traffic controller must talk with a train driver through a protocol that ensures a controlled safe passage.	Medium
7	Train traffic controller informs train driver	Multiple train drivers and colleagues need to be called and informed about the disruption.	Medium
8	Railroad crossing indicates malfunction	Multiple safety procedures need to be performed by the train traffic controller.	High

RESULTS

Perceived Obtrusiveness

The qualitative findings from the debriefing session showed that none of the participants perceived any of the sensors as being obtrusive. However, operators noted that the emWave was occasionally noticeable due to its wired connection with the computer. Most participants indicated that they forgot they were wearing the E3 wristband, as they were used to wearing watches. Few participants who were not used to wearing watches stated that they had sometimes found the E3 device noticeable. Overall, the BioPatch sensor was perceived as the least obtrusive device.

Usability of the Sensors for Research

The BioPatch sensor is provided with a heart rate confidence value as a quality indicator. Values >20% are indicated as a threshold for a reliable heart rate (Zephyr, 2014). For this study, this margin has been adopted based on the assumption that a FDA-approved sensor can validly serve as a gold standard device. A raw data file from the E3 can be extracted only from the website, which omits values when the

confidence rate of the measurement is not high enough. The emWave sensor always provides a value at each constant measurement point. Due to the unavailability of data points caused by a low measurement confidence, especially in the case of the E3, a normal distribution of RR intervals in each phase was not always obtainable. In the evaluation of usable data points, four data sets of the BioPatch showed a low heart rate confidence value, whereas seven data sets of the E3 partially showed too few data points to obtain a normal distribution. Also one data set from the E3 was lost during an attempt to upload the data from the device. No issues were encountered regarding the emWave.

The E3 and emWave data first needed to be processed manually by converting the time stamps and identifying the start of the recording. Time indications of the ECG file are directly displayed in Kubios. Some data files had to be manually split, since Kubios is not able to open very large files.

A precondition for the comparison among the three sensors is that the data in each phase have normal distributions and sufficiently high data confidence rates. Failure to meet this precondition led to omission from the analysis. Due to the omission or unavailability of data, only HRV data from three to six participants could be included in the analysis per phase.

Accuracy of the Sensors

Since the BioPatch is an FDA-approved ECG sensor, it was used as a reference for the accuracy of the E3 and emWave devices. For each participant, only full data sets from all three sensors were used in the analysis. Tables 4 and 5 present 10 HRV parameters and their means and standard deviations for each sensor and per phase, respectively for Scenario 1 (Phases 1–4) and Scenario 2 (Phases 6–8). One phase in Scenario 2 was removed, as the scenario was too short to provide sufficient data points for the E3 sensor.

To identify to what extent the E3 and emWave followed a similar trend in their data points in comparison to the BioPatch, which indicates the accuracy in relative terms, a linear regression was computed based on the data from the seven phases in Tables 4 and 5. The linear regression analysis was conducted between the BioPatch and each of the two other sensors; the significant relationships are given in Tables 4 and 5. In total, 6 of 10 HRV parameters showed a linear relationship between the emWave and the BioPatch data: mean RR, adjusted $R^2 = .99$, $F = 833.1$, $p < .001$; pNN50, adjusted $R^2 = .35$, $F = 4.3$, $p = .09$; LF power, adjusted $R^2 = .82$, $F = 27.5$, $p = .003$; normalized LF, adjusted $R^2 = .49$, $F = 6.8$, $p = .04$; normalized HF, adjusted $R^2 = .50$, $F = 6.9$, $p = .05$; and LF/HF ratio, adjusted $R^2 = .50$, $F = 7.1$, $p = .04$. A significant linear regression was found for 1 of 10 HRV parameters for the E3 sensor (i.e., mean RR, adjusted $R^2 = .98$, $F = 307.1$, $p < .001$).

The results from two commonly used HRV parameters derived from the time- and frequency-domain analysis (mean RR and normalized LF, respectively) are given in Tables 4 and 5.

Time-domain analysis. For the mean RR values (see Figure 4), a linear regression analysis was conducted over all phases. The findings show that the data from the BioPatch can be very well represented by a linear function in the emWave data (adjusted $R^2 = .99$, $F = 833.1$, $p < .001$, as listed earlier). The absolute values of the BioPatch and emWave can even be considered equal within the current measurement range, with an error of approximately 8%. For the E3, a similar result is observed (adjusted $R^2 = .98$, $F = 307.1$, $p < .001$, as also previously mentioned). Thus, the data values recorded by the emWave, E3, and BioPatch are linearly dependent and can easily be converted for the mean RR values as an HRV parameter.

Frequency-domain analysis. The regression analysis shows a poor correspondence for the normalized LF values between the emWave and the BioPatch (adjusted $R^2 = .49$, $F = 6.8$, $p = .04$) but a very good one when the outlier P7 is removed (adjusted $R^2 = .95$, $F = 99.1$, $p < .001$; see mean normalized LF values in Figure 5). For the E3 sensor, no such linear dependency is found. This implies that the findings from the E3 sensor do not provide accurate indications for the measurement of HRV for this parameter and should therefore not be used to derive conclusions.

TABLE 4: Heart Rate Variability Parameters for the E3, emWave, and BioPatch for Phases 1–4 (Mean ± *SD*)

	Phase 1 (n = 5)			Phase 2 (n = 6)			Phase 3 (n = 4)			Phase 4 (n = 3)		
HRV parameter	E3	eW	BP	E3	eW	BP	E3	eW	BP	E3	eW	BP
Mean RR,[a,b] ms	882.5 ± 126.6	856.7 ± 128.7	856.9 ± 131.9	880.0 ± 126.9	862.6 ± 125.6	864.8 ± 127.2	926.6 ± 154.8	897.6 ± 153.4	900.5 ± 153.4	808.5 ± 115.7	789.8 ± 125.0	798.7 ± 129.3
SDNN, ms	76.0 ± 17.1	60.3 ± 14.8	45.9 ± 16.2	74.1 ± 21.7	53.7 ± 21.5	39.9 ± 6.0	89.6 ± 29.5	66.1 ± 23.5	42.5 ± 12.6	70.1 ± 7.9	56.4 ± 13.4	43.9 ± 4.2
RMSSD, ms	91.0 ± 23.7	57.8 ± 18.9	33.7 ± 11.0	71.0 ± 20.3	53.4 ± 16.2	34.0 ± 11.2	86.0 ± 33.8	64.7 ± 18.0	32.7 ± 10.3	77.7 ± 15.3	61.3 ± 24.8	31.3 ± 7.1
pNN50,[c] %	45.9 ± 12.9	29.4 ± 8.5	10.2 ± 9.0	30.5 ± 13.8	25.8 ± 9.7	8.4 ± 5.3	37.3 ± 14.8	33.0 ± 8.0	12.6 ± 10.8	41.0 ± 10.0	34.0 ± 17.7	12.0 ± 9.0
TINN, ms	265.0 ± 102.0	301.0 ± 48.7	270.0 ± 93.6	266.7 ± 111.3	285.0 ± 70.5	239.2 ± 95.8	222.5 ± 99.9	336.3 ± 83.7	201.3 ± 52.8	260.0 ± 80.0	265.0 ± 62.4	205.0 ± 5.0
LF power,[a] ms^2	11,118.6 ± 11,135.4	1,688.2 ± 1,103.8	1,633.2 ± 1,412.4	6,891.0 ± 5,364.0	1,057.2 ± 364.9	1,068.3 ± 403.9	6,772.8 ± 5,595.2	1,384.5 ± 432.1	1,513.8 ± 756.8	13,326.7 ± 15,995.9	1,646.0 ± 230.2	1,497.7 ± 337.6
HF power, ms^2	1,065.6 ± 777.8	702.0 ± 411.6	276.2 ± 104.0	578.2 ± 407.0	654.8 ± 383.8	338.0 ± 162.1	472.8 ± 230.4	946.3 ± 752.1	419.3 ± 136.0	855.0 ± 320.2	1,224.3 ± 1,150.4	477.7 ± 132.1
LF, n.u.[a]	86.5 ± 12.8	67.5 ± 13.4	82.0 ± 7.4	87.4 ± 9.2	62.1 ± 13.4	74.8 ± 13.1	83.1 ± 23.1	62.4 ± 14.7	76.4 ± 8.5	86.8 ± 12.0	62.5 ± 22.1	75.2 ± 8.8
HF, n.u.[a]	13.5 ± 12.8	32.3 ± 13.4	18.0 ± 7.4	12.6 ± 9.2	37.8 ± 13.1	25.2 ± 13.1	16.9 ± 23.0	37.5 ± 14.6	23.6 ± 8.5	13.1 ± 11.9	37.4 ± 22.0	24.7 ± 8.8
LF/HF ratio[a]	11,133.4 ± 7,708.6	2,997.2 ± 2,862.3	5,727.2 ± 3,737.0	44,678.5 ± 89,766.1	1,848.0 ± 1,366.6	4,001.5 ± 2,561.0	13,518.7 ± 9,447.8	1,746.2 ± 1,216.1	3,594.5 ± 1,382.4	12,841.7 ± 11,799.0	2,270.5 ± 2,487.0	3,395.7 ± 1,521.7

Note. HRV, heart rate variability; eW, emWave; BP, BioPatch; RR, difference between R-wave occurrence times in milliseconds; SDNN, standard deviation of normal-to-normal RR intervals; RMSSD, root mean square of successive differences in milliseconds; pNN50c, relative amount of successive intervals differing >50 ms in percentages; TINN, triangular interpolation in milliseconds; LF, lower frequency; HF, higher frequency; n.u., normalized units.

[a]eW and BP, $p \leq .05$. [b]E3 and BP, $p \leq .05$. [c]eW and BP, $p \leq .10$.

TABLE 5: Heart Rate Variability Parameters for the E3, emWave, and BioPatch for Phases 6–8 (Mean ± *SD*)

	Phase 6 (n = 4)			Phase 7 (n = 5)			Phase 8 (n = 3)		
HRV parameter	E3	eW	BP	E3	eW	BP	E3	eW	BP
Mean RR,[a,b] ms	811.1 ± 49.1	792.8 ±41.9	793.1 ± 43.6	804.8 ± 41.1	801.7 ± 42.9	801.1 ± 44.4	880.2 ± 40.0	854.0 ± 36.9	854.8 ± 32.6
SDNN, ms	75.9 ± 3.2	55.9 ± 12.7	44.5 ± 12.7	68.4 ± 8.8	58.3 ± 10.3	43.3 ± 10.1	61.0 ± 11.4	59.1 ± 6.9	42.1 ± 4.4
RMSSD, ms	84.3 ± 15.2	57.8 ± 18.2	31.3 ± 6.8	76.6 ± 17.9	63.8 ± 16.2	31.9 ± 9.4	75.6 ± 21.9	70.7 ± 8.7	31.1 ± 4.8
pNN50,[c] %	45.5 ± 7.5	29.0 ± 11.1	8.5 ± 3.0	35.1 ± 9.8	29.6 ± 9.4	8.2 ± 4.7	37.4 ± 7.0	34.6 ± 2.6	9.6 ± 2.2
TINN, ms	320.0 ± 37.0	287.5 ± 68.9	221.3 ± 68.0	316.0 ± 74.4	337.0 ± 67.1	254.0 ± 75.1	220.0 ± 79.4	295.0 ± 37.7	223.3 ± 45.4
LF power,[a] ms^2	8,329.0 ± 5,749.2	1,663.5 ± 919.2	1,695.5 ± 1 ,258.0	7,604.4 ± 3,943.8	1,561.8 ± 746.0	1,463.6 ± 756.7	6,335.7 ± 5,480.2	1,269.0 ± 402.1	1,183.7 ± 594.6
HF power, ms^2	494.5 ± 137.4	988.3 ± 470.5	498.3 ± 229.2	528.2 ± 231.9	921.0 ± 509.6	713.2 ± 1,012.0	495.3 ± 376.9	1,399.0 ± 398.4	407.7 ± 208.0
LF, n.u.[a]	91.1 ± 9.0	62.0 ± 12.3	75.3 ± 6.6	92.9 ± 1.8	62.8 ± 16.1	75.5 ± 17.1	92.2 ± 0.9	47.3 ± 14.6	72.4 ± 13.2
HF, n.u.[a]	8.9 ± 8.9	37.9 ± 12.2	24.7 ± 6.6	7.1 ± 1.9	36.9 ± 16.0	24.4 ± 17.1	7.8 ± 0.9	52.3 ± 14.9	27.5 ± 13.2
LF/HF ratio[a]	17,036.5 ± 9,230.1	1,630.7 ± 1,241.0	3,270.5 ± 1,089.1	13,717.0 ± 2,862.0	1,922.9 ± 1,187.0	4,415.2 ± 3,286.5	12,013.0 ± 1,566.7	854.5 ± 761.3	3,344.0 ± 2,320.8

Note. HRV, heart rate variability; eW, emWave; BP, BioPatch; RR, difference between R-wave occurrence times in milliseconds; SDNN, standard deviation of normal-to-normal RR intervals; RMSSD, root mean square of successive differences in milliseconds; pNN50c, relative amount of successive intervals differing >50 ms in percentages; TINN, triangular interpolation in milliseconds; LF, lower frequency; HF, higher frequency; n.u., normalized units.
[a]eW and BP, $p \leq .05$. [b]E3 and BP, $p \leq .05$. [c]eW and BP, $p \leq .10$.

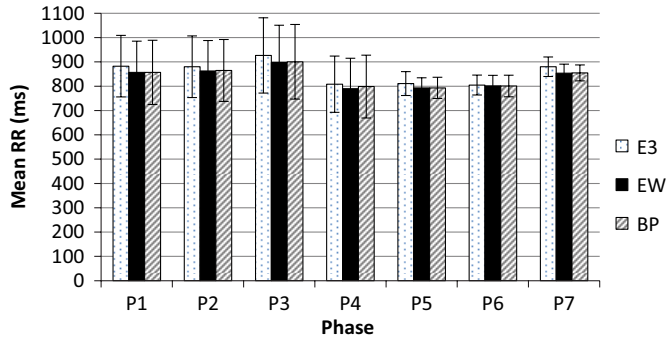

Figure 4. Mean difference between R-wave occurrence times (RR) with standard deviations of the E3, emWave (eW), and BioPatch (BP) in each phase (P).

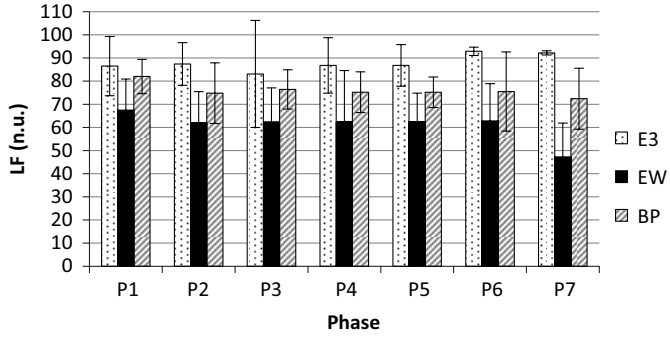

Figure 5. Normalized lower-frequency (LF) values with standard deviations of the E3, emWave (eW), and BioPatch (BP) in each phase (P).

DISCUSSION AND CONCLUSION

In the present study, three psychophysiological sensors were investigated on their perceived obtrusiveness, their usability for research purposes in terms of usable data points, and the ease of postprocessing data and accuracy on HRV (the E3 and emWave vs. the FDA-approved BioPatch as a gold standard reference device).

The data points of the E3, the emWave, and the BioPatch are linearly dependent when the analysis of HRV is limited to the mean RR parameter. For this HRV parameter, the absolute values of the emWave could be used interchangeably with those of the BioPatch. However, when the results from all 10 HRV parameters were taken into account, a correspondence was found for 6 of 10 parameters between the emWave and the BioPatch. For the E3 sensor, this applies to only one HRV parameter. Thus, great care should be taken when selecting HRV parameters derived from the E3 wristband data.

These findings are surprising in that, overall, the emWave is more accurate than the E3, given the lower sampling rate of the EmWave. A reason for the difference in significant findings between the E3 and the emWave with regard to the 10 HRV parameters might be the location of the sensors: A wristband sensor might incur more noise during data acquisition. The impact of the E3 as being less accurate given its significant relationship with 1 of 10 HRV parameters may contribute to the discussion about the extent to which commercially available wearable sensors, often wristbands, are accurate. However, although the FDA does not actively enforce regulations on commercially available wearable sensors, it has developed a draft set of guidelines for products that are used only for wellness purposes, that present a very low risk to users'

safety, and for which the data are not interpreted to diagnose diseases or conditions (U.S. Department of Health and Human Services, 2015).

The participants did not perceive any of the three sensors as obtrusive, although a few found the emWave and the E3 noticeable at times. In terms of usable data points, both the E3 and the BioPatch had difficulty acquiring either sufficiently high data-recording confidence values (especially the BioPatch) or normal distributions (especially the E3, due to missing data points in the output file). As for the ease of postprocessing data, the BioPatch output files required no manual postprocessing, which reduced costs and efforts in the analysis stage. In this case, 12 participants were analyzed, and in the case of both the emWave and the E3, data had to be prepared for the actual analysis in terms of manually identifying the start and end phases within the output files (e.g., through time stamp conversions, the synchronization of sampling rate with times, and adaption of the output file). The postprocessing of data took a considerable effort over a period of a few days for both files. An output file that needs very few or no manual adaptations can be highly recommended when dealing with a larger sample. However, greater compatibility between the output file and the software program can also support the postprocessing work.

Each of the three devices has its own pros and cons regarding deployment in a research setting. The BioPatch would be the overall preferred sensor, due to its assumed accuracy and usability for research purposes. However, a strong emphasis lies on exploring how to increase the data-recording confidence rate. As an alternative sensor, the emWave could be acceptable for measuring HRV, as it had a sufficiently high accuracy on 6 of 10 parameters to measure HRV, in which 1 parameter (mean RR) was significant in terms of absolute values. It is not, however, a wearable sensor, which limits the measurements to a static setting. Note that sensors can be more preferred or recommended given the priorities of a particular study. For instance, if the number of data points is prioritized, if no time constraints are set on postprocessing data, and if no ambulatory movements are expected during the study, the emWave would be recommended over the BioPatch, as the latter's data-recording confidence rate is unstable. The E3 could also be used in control room studies when researchers want to use a wristband sensor and are supportive of using only the mean RR as an HRV parameter to calculate MWL.

This study shows that the wearable BioPatch sensor and the low-cost emWave sensor can both be used to analyze the MWL development of train traffic controllers. The positive advantages of the sensors, especially with regard to the ease of use in recording and postprocessing data for analysis, could provide opportunities for cost reductions in terms of expertise in traditional ECG measurements as well as in equipment costs. With proper but limited training, it is possible to apply the devices and prepare the data set without much prior experience. Moreover, with the increasing popularity of wearable sensors, the chance increases that researchers and participants are acquainted with the use of these sensors. Additionally, in the light of so-called regret costs, measuring the objective MWL of operators through low-cost heart rate sensors (<$450 for the BioPatch sensor [in U.S. dollars]) costs a fraction of the amount spent on implementing a traffic management system—for example, approximately $43 million (originally stated as 28 million pounds) in the United Kingdom (*Railway Gazette*, 2014).

However, limitations in this study should be noted, as the BioPatch had a different sensor placement on the chest on the second measurement day. As the sample size was too small to investigate possible differences between the two groups, future research should investigate possible effects. In general, future research is needed to confirm these findings, given the limited number of samples in this study. Also the presented conclusions are valid only for HRV measurements in nonphysically demanding tasks, such as supervisory control tasks. More research is needed to explore the other wearable sensors on the market, to investigate the effects of individual differences, such as gender and age, on the level of accuracy of the sensors and to confirm the use of the BioPatch as a valid and accurate reference device.

All in all, the accessibility and affordability of wearable sensors show great potential to obtain insights in the MWL development of

operators next to the use of observational and/or self-rating data. The advantages of long-term measurements throughout operators' shifts can especially provide new insights into the cognitive demands of certain tasks at hand. Future work could also examine the MWL through HRV and stress through EDA relation, which can be retrieved from the Empatica device.

ACKNOWLEDGMENTS

This research was funded through the Railway Gaming Suite program, a joint project of ProRail and Delft University of Technology. We thank Berend Wouda and Giel van Lankveld from Delft University of Technology and Rolf Zon from the National Aerospace Laboratory for their support and collaboration in this study. We are also grateful for the discussions on the sensors with Professor Raja Parasuraman from George Mason University, who passed away last year. Thank you for your inspiring work.

REFERENCES

Brookhuis, K. A. (2004). Psychophysiological methods. In N. A. Stanton, A. Hedge, K. A. Brookhuis, E. Salas, & H. W. Hendrick (Eds.), *Handbook of human factors and ergonomics methods*. Boca Raton, FL: CRC Press.

Brookhuis, K. A., & de Waard, D. (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis & Prevention*, *42*(3), 898–903.

Cain, B. (2007). *A review of the mental workload literature*. Toronto, Canada: Defence Research and Development Toronto.

Empatica. (2013). *E3 User Manual 2013-09-15*. Milan, Italy: Author.

Essner, A., Sjöström, R., Ahlgren, E., & Lindmark, B. (2013). Validity and reliability of Polar® RS800CX heart rate monitor. Measuring heart rate in dogs during standing position and at trot on a treadmill. *Physiology & Behavior*, *114–115*, 1–5.

Fagundes, C. P., Murray, D. M., Hwang, B. S., Gouin, J.-P., Thayer, J. F., Sollers, J., III, . . . Kiecolt-Glaser, J. K. (2011). Sympathetic and parasympathetic activity in cancer-related fatigue: More evidence for a physiological substrate in cancer survivors. *Psychoneuroendocrinology*, *36*(8), 1137–1147.

Garbarino, M., Lai, M., Bender, D., Picard, R. W., & Tognetti, S. (2014, November). *Empatica E3—A wearable wireless multisensor device for real-time computerized biofeedback and data acquisition*. Paper presented at the Fourth International Conference on Wireless Mobile Communication and Healthcare, Athens, Greece.

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

Hlotova, Y., Cats, O., & Meijer, S. (2014). Measuring bus drivers' occupational stress under changing working conditions. *Transportation Research Record: Journal of the Transportation Research Board*, *2415*, 13–20.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.

Lo, J. C., & Meijer, S. A. (2013). Measuring group situation awareness in a multi-actor gaming simulation: A pilot study of railway and passenger traffic operators. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Lo, J. C., Van den Hoogen, J., & Meijer, S. A. (2013). Using gaming simulation experiments to test railway innovations: Implications for validity. In *Proceedings of the 2013 Winter Simulation Conference* (pp. 1766–1777). New York, NY: IEEE.

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Julian Abich, I. (2015). The psychometrics of mental workload. *Human Factors*, *57*(1), 125–143.

Meijer, S. A. (2012). Introducing gaming simulation in the Dutch railways. *Procedia: Social and Behavioral Sciences*, *48*, 41–51.

Mulder, L. J. M., De Waard, D., & Brookhuis, K. A. (2004). Estimating mental effort using heart rate and heart rate variability. In N. A. Stanton, A. Hedge, K. A. Brookhuis, E. Salas, & H. W. Hendrick (Eds.), *Handbook of human factors and ergonomics methods*. Boca Raton, FL: CRC Press.

Parasuraman, R. (2011). Neuroergonomics: Brain, cognition, and performance at work. *Current Directions in Psychological Science*, *20*(3), 181–186.

Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, *9*(1), 1–17.

Pickup, L., Wilson, J. R., Sharples, S., Norris, B., Clarke, T., & Young, M. S. (2005). Fundamental examinations of mental workload in the rail industry. *Theoretical Issues in Ergonomics Science*, *6*(6), 463–482.

Pina, P. E., Donmez, B., & Cummings, M. L. (2008). *Selecting metrics to evaluate human supervisory control applications*. Cambridge, MA: MIT Humans and Automation Laboratory.

Porto, L. G. G., & Junqueira, L. F., Jr. (2009). Comparison of time-domain short-term heart interval variability analysis using a wrist-worn heart rate monitor and the conventional electrocardiogram. *Pacing and Clinical Electrophysiology*, *32*(1), 43–51.

ProRail. (2015). *Jaarverslag 2014* [Annual report 2014]. Retrieved from https://www.prorail.nl/sites/default/files/jaarverslag_2014.pdf

Railway Gazette. (2014). Thales wins network rail traffic management system contracts. Retrieved from http://www.railwaygazette.com/news/infrastructure/single-view/view/thales-wins-network-rail-traffic-management-system-contracts.html

Reid, G. B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A scaling procedure for measuring mental workload. *Advances in Psychology*, *52*, 185–218.

Schäfer, A., & Vagedes, J. (2013). How accurate is pulse rate variability as an estimate of heart rate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram. *International Journal of Cardiology*, *166*(1), 15–29.

Tarvainen, M. P., Niskanen, J.-P., Lipponen, J. A., Ranta-aho, P. O., & Karjalainen, P. A. (2014). Kubios HRV—Heart rate variability analysis software. *Computer Methods and Programs in Biomedicine*, *113*(1), 210–220.

U.S. Department of Health and Human Services. (2015). *General wellness: Policy for low risk devices 2015-01-20*. Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM429674.pdf?source=govdelivery&utm_medium=email&utm_source=govdelivery

Vidulich, M. A., & Tsang, P. S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. *Proceedings of the Human Factors Society*, *31*(9), 1057–1061.

Wallén, M., Hasson, D., Theorell, T., Canlon, B., & Osika, W. (2012). Possibilities and limitations of the polar RS800 in measuring heart rate variability at rest. *European Journal of Applied Physiology*, *112*(3), 1153–1165.

Weippert, M., Kumar, M., Kreuzfeld, S., Arndt, D., Rieger, A., & Stoll, R. (2010). Comparison of three mobile devices for measuring R-R intervals and heart rate variability: Polar S810i, Suunto t6 and an ambulatory ECG system. *European Journal of Applied Physiology*, *109*(4), 779–786.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement applications. *Proceedings of the Human Factors Society*, *27*(2), 129–133.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, *35*(2), 263–281.

Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *International Journal of Aviation Psychology*, *12*(1), 3–18.

Wilson, J. R., & Norris, B. J. (2005). Rail human factors: Past, present and future. *Applied Ergonomics*, *36*, 649–660.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2014). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17.

Zephyr. (2013). *BioPatch Data Sheet 2013-06-14*. Annapolis, MD: Medtronic.

Zephyr. (2014). *BioHarness 3: Log Data Descriptions 2014-04-04*. Annapolis, MD: Medtronic.

Zijlstra, F. R. H. (1985). *The construction of a scale to measure perceived effort*. Delft, Netherlands: Delft University Press.

Julia C. Lo is a human factors project manager at ProRail and a PhD candidate at the Faculty of Technology, Policy, and Management of Delft University of Technology. Areas of interest include multidisciplinary research from cognitive, social, and organizational sciences in operational environments. Her PhD research focuses on the investigation of individual and team situation awareness in the railway sector using gaming simulations as a research method.

Emdzad Sehic is a project manager at the innovation and development department at ProRail and an external PhD candidate at the Faculty of Technology, Policy, and Management of Delft University of Technology. His main research interest is the development of a distributed gaming simulation methodology for decision making within the railway sector.

Sebastiaan A. Meijer is a professor in the School of Technology and Health at KTH Royal Institute of Technology and is associated with the Faculty of Technology, Policy, and Management at Delft University of Technology. He specializes in gaming simulation and other interactive methods to involve the operational level of organizations in innovation processes.