What to Put on the User: Sensing Technologies for Studies and Physiology Aware Systems

Katrin Hänsel
Queen Mary University of London
k.hansel@qmul.ac.uk

Romina Poguntke
University of Stuttgart
romina.poguntke@vis.uni-stuttgart.de

Hamed Haddadi
Imperial College, London
h.haddadi@imperial.ac.uk

Akram Alomainy
Queen Mary University of London
a.alomainy@qmul.ac.uk

Albrecht Schmidt
Ludwig Maximilian University, Munich
albrecht.schmidt@um.ifi.lmu.de

ABSTRACT

Fitness trackers not just provide easy means to acquire physiological data in real-world environments due to affordable sensing technologies, they further offer opportunities for physiology-aware applications and studies in HCI; however, their performance is not well understood. In this paper, we report findings on the quality of 3 sensing technologies: PPG-based wrist trackers (Apple Watch, Microsoft Band 2), an ECG-belt (Polar H7) and reference device with stick-on ECG electrodes (Nexus 10). We collected physiological (heart rate, electrodermal activity, skin temperature) and subjective data from 21 participants performing combinations of physical activity and stressful tasks. Our empirical research indicates that wrist devices provide a good sensing performance in stationary settings. However, they lack accuracy when participants are mobile or if tasks require physical activity. Based on our findings, we suggest a *Design Space for Wearables in Research Settings* and reflected on the appropriateness of the investigated technologies in research contexts.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Wearable Technology, Validation, Stress, Affective Computing

INTRODUCTION

Recent advances in consumer wearables allow the ubiquitous collection of health data, such as physical activity or sleep in everyday life. Enabled through the variety of affordable consumer wearables flooding the market[1] every year, wire-free and independent tools for assessing physiological data are becoming increasingly valuable for researchers and scientists alike. With the feasibility to measure signals (e.g. heart rate response) 'on the fly' and in daily life situations, there come along immense opportunities.

Rosalind Picard was among the first to emphasize the importance of sensing wearables. In the article "Affective Wearables" [48], she discussed application scenarios and presented a prototype for recording physiological data, like blood volume pressure, Galvanic Skin Response and respiration. In the past years, various systems exploiting the feasibility to access physiological data have been published approaching more and more user-adaptive interfaces [6] and systems [57]. Emerging fields in HCI, like calm-computing [38] and avoidance of technostress [74], can benefit from ubiquitous, wearable affect and stress sensing technology and adaptive systems.

However, much of the understanding of physiological sensing is based on high accuracy lab equipment and it remains unclear how well consumer devices are suited for this purpose. Khusainov et al. [36] discuss in their survey paper that wearable sensors often lack accuracy and appropriate sampling rates. Furthermore, not every wearable system delivers raw data; e.g. the Fitbit and Jawbone[2] device families do not allow users to assess their raw physiological data. Consequently, there is a need to evaluate the reliability of wearable consumer technology with regards to their accuracy and suitability for physiological and psychological research applications.

In this paper, we perform a comparison between two wrist-worn devices with optical heart rate sensors (Apple Watch, Microsoft Band 2[3]) against a heart rate chest strap (Polar H7[4]) and a laboratory measurement instrument (Nexus 10 kit) under different physical and stressful conditions. To evaluate the

CHI 2018, April 21–26, 2018, Montreal, QC, Canada

ISBN 978-1-4503-5620-6/18/04...$15.00

DOI: https://doi.org/10.1145/3173574.3173719

[1] The market value for wearable devices is forecasted to be almost 6 billion USD by 2018 [32].

[2] www.fitbit.com and www.jawbone.com

[3] www.apple.com/watch and www.microsoft.com/microsoft-band

[4] www.polar.com

appropriateness of the aforementioned wearables, we perform the following research activities:

1. Comparing accuracy in physiological data measured by the different wearable technologies
2. Examining how physiologically measured stress is affected under stationary and physical activity
3. Investigating correlations between subjective measures and physiological data
4. Introducing a *Design Space for Measurement Tools*, reflecting five dimensions which are important to consider for the choice of measurement technology in research contexts

RELATED WORK

Our research addresses the feasibility of consumer smart wearables for research settings with a focus on stress assessment. We therefore provide background information on physiological and subjective perceived measures to detect stress and how wearables can provide sufficient data in this field.

Physiological Data and Stress

The human body is a complicated and continuously working system. We receive many physiological responses indicating stress; we breath faster, blood pressure and pulse rate increase, and we begin to sweat, to name only a few indicators. These reactions are attributable to the activation of the sympathetic nervous system which autonomously triggers a series of physiological changes [13, 64]. Those changes can be picked up by sensors to make stress predictions.

The heart rate signal, as an indicator of physiological changes, has been used as in various studies among different disciplines such as medicine [26], psychology [22, 59], and HCI [43] due to its sufficient reliability and data richness. *Electrodermal activity* (EDA), which is mostly referred to as the activation of sweat glands and hence can be called *Galvanic Skin Response* (GSR) or *skin conductivity* [14], can be found in prior work as an indicator for cognitive load [62], stress [29] and also as a "predictor of emotional responses to stressful life events" [46]. As a third measure, we choose skin temperature due to its good prediction ability indicating stress [35] through significant changes in body temperature [71].

Assessing Physiological Data through Wearables

Recently, fitness trackers and smart watches became increasingly popular and ubiquitous. While early devices focused merely on activity tracking via step count or flights of stairs climbed, modern devices incorporated additional biophysiological sensors such as Photoplethysmography (PPG), skin conductance or skin temperature sensors to provide a fuller picture of the consumers fitness and health patterns. Various device manufacturers provide devices with closed systems and proprietary algorithms for the estimation of physical activity, heart rate or energy expenditure, but evidence for the validity and reliability of the provided health data is sparse for the variety of devices.

Electrocardiography (ECG) is the process of recording the electrical activity of the heart and it is widely used to extract the heart rate with electrodes placed on the chest and by detecting peaks in the signal. On the contrary, consumer devices commonly rely on using optical Photoplethysmography (PPG) sensors to extract heart rate from peaks in the blood flow under the skin [68]; this happens predominantly on the wrist. Depending on the placement of the optical sensor, there is a time delay in the detected peak in blood flow caused by a heart beat called Pulse Transit Time (PTT). This can potentially lead to errors in beat detectiion; still, various studies compared both technologies for their ability to detect heart beats and heart beat intervals and found good correlation in the ECG gold standard and PPG [41, 49, 60].

Instead of considering the beat-per-beat detection, various studies focused on comparing the reported heart rate; there is evidence that PPG devices show a decline in accuracy compared to the gold standard with increased physical activity and heart rate [34, 47, 68]. Further, Spierer et al. [65] found particular differences in heart rate agreeability depending on skin pigmentation between the two wrist-worn devices Mio Alpha and Omron HR500U; while sensitive skin types (Type II on Fitzpatrick Scale [23]) showed similar low error rates, the error rates significantly increased for the Mio Alpha (for skin type V). These findings highlight a manufacturer dependent variation in accuracy.

Subjective Measures as Indicators for Stress

Apart from the physiological indicators of stress based on the sympathetic nervous system's reactions, there are self-rating measures to assess stress. While in the beginning of the research established around stress, subjective assessment methods alone were used [12], later questionnaires have been applied as ground truth measures to compare against other measures, i.e. physiological sensors. Kramer [37] argues that physiological sensors captures changes that can be monitored within seconds whereas subjective rating of one's stress level only provide snapshots. On the contrary, subjective assessment are tools easy to operate for participants and experimenters.

One tool to assess affective states is the Self-Assessment Manikin (SAM) [11]. This tool quickly and reliably collects the participants' perception of their moods on three dimensions: arousal, valence and dominance. Its values have been shown to match emotional and stress responses [24, 51] and the responses were found to be cross-cultural observable [45].

COMPARING DIFFERENT SENSING TECHNOLOGIES

In this work, we aim to validate devices with 3 different heart rate sensing technologies for their ability to infer stress and increased arousal in a controlled lab environment: two consumer wearables with optical heart rate technologies (Apple Watch Series 2 and Microsoft Band 2), an ECG-belt device (Polar H7) and a laboratory measurement instrument with ECG adhesive electrodes (Nexus kit 10). In the following, we will present the underlying concept of our work explaining our choice of our wearables and measures, further deducing our hypotheses from related literature.

Choice of Physiological Stress-indicating Measures

Several studies used physiological measures i.e. heart rate, electrodermal activity, and skin temperature to detect stress and showed correlation with subjective stress responses [56].

Moreover, the combination of these measures has proved to be a reliable indicator in e.g. psychology [2, 4], for the development of a non-invasive real-time stress tracking system [39], in a real-world driving tasks to determine the driver's stress level [28], or for non-invasive stress detection in HCI [5]. This becomes increasingly interesting with respect to future works.

Choice of Sensing Technologies

When it comes to measuring heart rate, there are two prevalent technologies: Photoplethysmography (PPG) and Electrocardiography (ECG). Most wrist-based consumer devices rely on optical heart rate sensing with PPG sensors; however, the research gold standard is ECG [34] whereby heart beats are detected via the electrical signal-signature of the heart. Electrodes can hereby either be self-adhesive and stick-on or held on place by an elastic chest strap. With focus on heart rate, we picked the following devices for these sensing technology categories.

The Apple Watch, as a popular smartwatch with fitness capabilities in form of physical activity and heart rate tracking. In several studies, the Apple Watch performed best compared to other wrist devices in terms of heart rate error and correlation with the gold standard device [16, 63, 73]. The Microsoft Band 2 fitness tracker, as another optical heart rate device, has been chosen for its rich sensor set and accessibility of data. It is one of the few consumer wearables incorporating skin conductance and skin temperature sensors.

Contrary to the often used stick-on ECG-electrodes used in medical and laboratory settings, chest-belt heart rate monitors can be used without the need for adhesives due to the electrodes being held in place by an elastic strap. This technology has been shown to have a high accuracy [25] and have been used as criterion devices in related work [66]. We chose the Polar H7 chest strap as an exemplary device for our study based on its ability to share sensing data via Bluetooth to a mobile phone.

The chosen wearables provide programming interfaces for the iOS environment which was leveraged to build a proprietary app for the data collection, aggregation and synchronization to provide the necessary data for our study purpose.

As a laboratory measurement instrument, we use the Nexus-10 MK2 by Mind Media[5]. This is a wireless device which is targeted for biofeedback applications and psychological research. It offers a range of channels for various sensors. In this study, we utilized the ECG signal through self-adhesive electrodes (Lead II setup, as instructed in the device manual), GSR finger electrodes, and skin temperature sensor placed at the participants' forearm. The manufacturers BioTrace+ software allows real-time data visualizations, recording and marker placement functionalities.

Hypotheses

In this work, we address the devices' ability to identify differences in physiological data during relaxed and stressful situations and how physical activity affects the measurements recorded by consumer devices. Further, we investigated correlations between arousal as a stress indicator and physiological data.

Based on previous research, we hypothesized that there will be a lessened accuracy and correlation of the wrist-devices devices in physical activity compared to stationary activity, i.e. a difference in heart rate recorded via ECG and PPG. The PPG signal, which is used in the wrist-worn devices Apple Watch and Microsoft Band, is prone to movement artifacts [1]. Validation studies such as Tamura et al. [68] confirmed the decreased accuracy of wrist-measured heart rate in consumer devices. Therefore, we phrased our hypotheses as follows:

H 1a There is a difference in the physiological data measured by different devices under physical activity

H 1b There is no difference in the physiological data measured by different devices under stationary activity

Further, there is related work [22, 43, 62], indicating that physiological and subjective measures differ in relaxed compared to stressed states. According to our experimental design, we added the dimension of physical activity. This enabled us to verify the aforementioned finding with respect to physical activity, hypothesizing:

H 2a There is a difference in physiological data between stressful and relaxed situation under physical activity

H 2b There is a difference in physiological data between stressful and relaxed situation under stationary activity

Lastly, we focused on the relation between subjectively perceived measures and physiological data. The subjective measures arousal, valence and dominance were hereby assessed with the Self-Assessment Manikin [11]. This was accompanied by the additional assessment of 'awakeness' and 'tension' as argued by former work [15, 58].

Prior work from neuropsychology suggested that there are correlations among neurobiological processes triggering the increase of stress hormones and perceived stress [24, 55], arousal [24, 51], and valence [24, 51]. Other studies found that heart rate activity increased when arousal and valence were higher [76]. Salimpoor et al. [54] showed that arousal and valence strongly correlated with electrodermal activity, body temperature, heart and respiration rate as well as blood volume pulse. Remarkably, dominance was not found to be correlating with an increase of stress hormones [51]. Due to these results, we aimed to investigate the following hypotheses:

H 3a There is a correlation in between stress perception and physiological data

H 3b There is a correlation between arousal and physiological data

H 3c There is a correlation between valence and physiological data

H 3d There is no correlation between dominance and physiological data

We answered these hypotheses by conducting a user study involving four trials combining activity and stress, which we will describe in the following section.

[5] www.mindmedia.info/CMS2014/en/products/systems/nexus-10-mkii

USER STUDY

In the following we, describe the measures of our experiment, the conditions and tasks we used in our study design, as well as the procedure and demography of our participants.

Study Design

For this study, we chose a within-subject design implying that each participant underwent all of our four conditions lasting 20 minutes in total (5 minutes per condition). We randomized the sequence of conditions according to Latin Square. Each condition was a combination of the two levels for each of our two independent variables, namely physical activity and stress. These two levels for physical activity consisted of walking on a treadmill and being seated, whereas stress was split into performing mental arithmetic tasks (MAT) and relaxing while listening to meditation music. Hence and by using factorial design, the conditions *relaxed walking (RW)*, *relaxed stationary (RS)*, *MAT walking (MW)*, and *MAT stationary (MS)* resulted (see Figure 1). A similar setup of conditions has been used by Sun et al. [67].

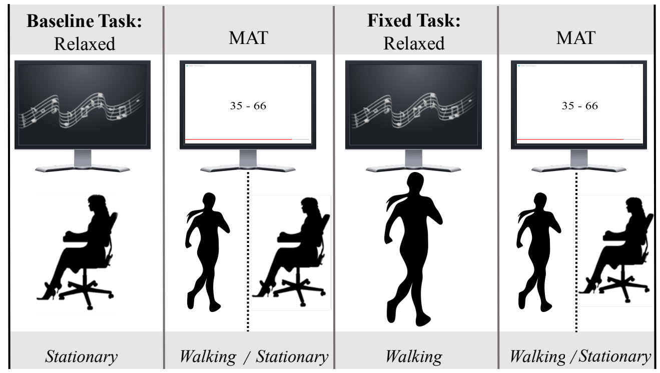

Figure 1: The figure shows the study overview and depicts the sequence of the trials according to our study design. It consisted of four trials for each participant including the baseline task in the beginning and the fixed task in the middle. In the second and the third trial we switched between the *walking* and *stationary* condition in counterbalanced order.

Independent Variables

Measurement Devices

For our study, we focused on two wrist-based consumer wearables (Apple Watch Series 2, Microsoft Band 2) equipped with physiological sensors, one chest strap heart rate monitor (Polar H7 chest belt), and a laboratory measurement instrument (Nexus 10 kit) serving as independent variables.

Physical Activity and Stress

Further, physical activity and stress served as our independent variable. Stress was divided into either performing mental arithmetic tasks or relaxing while listening to meditation music. Differentiating physical activity, we asked participants to either walk on a treadmill in their own, physiologically demanding pace or to remain stationary on a comfortable chair.

Dependent Variables

Physiological Data

As dependent variables, we recorded physiological data, namely heart rate, EDA and skin temperature, from the aforementioned devices. As discussed previously, these measures have been shown to provide high reliability indicating stress [2, 35, 62].

Self-Reported Arousal, Valence, and Dominance

For the self-reported measures for stress and affective state, we applied the widely-used Self-Assessment Manikin Scale (SAM) [11]. This scale allows the non-verbal assessment of current affective state, respectively valence (pleasure), arousal and dominance, through pictures. As in the original work by Bradley and Lang [11], we utilized a 9-point rating scale for each dimension whereby participants were instructed to place a 'x' on any of the five figures or between two figures.

The classical arousal dimension in this and similar models, e.g., Russell's Circumplex Model of Affect [53], does not differentiate between experienced tension; but based on Thayer [69], arousal can be further characterized by energetic arousal (ranging from wide-awake to tired) and tense arousal (nervous to calm). According to the recommendation of [58], we added two additional questions: a 5-point self-rating Likert-item for each dimension assessing tension and wakefulness [30, 52].

Self-Reported Stress

For the assessment of how stressful the task has been perceived, we used a single 5-point Likert scale ranging from 1(="not at all stressful") to 5(="very much stressful") [21, 28].

Tasks and Stimulus Material

Participants were asked to relax while listening to meditation music[6] and to perform mental arithmetic operations. As stimulus material, we presented mental arithmetic tasks for five minutes on a 60-inch display placed right in front of the participants. This task has been proven to induce stress [8] and to affect physiological parameters [27, 42, 61, 70]. The calculations, addition and subtraction of two-digit numbers ranging from 0-100 and including negative solutions, had be completed within 6 seconds each. A timeline signifying the time left for each task was displayed on the screen. Correct answers were rewarded with a green screen displaying "Correct". For false answers or when the time was up, participants heard a buzz sound and the screen displayed "False" or "time over" on red background. The visual countdown and feedback (both visual and auditory) had been proven to increase subjectively perceived and physiological stress [62]. Our study setup was inspired by Vlemincx et al. [72]. To perform the walking task, we asked participants to walk for five minutes on a treadmill (model: ProFitness Sierra motorized).

Participants and Procedure

For our laboratory study, we recruited 24 participants including one pilot test person via university mailing lists, leaflets and personal recruitment campaigns. Two participants and the pilot were excluded from data analysis due to technical problems during the data acquisition. The mean age of the

[6]As meditation music we used song number 14 from the album '72 Ambient Meditations'

		Heart Rate				Skin Temperature		EDA	
		Nexus	Polar	Microsoft Band	Apple Watch	Nexus	Microsoft Band	Nexus	Microsoft Band
RS	Mean	66.16	66.89	66.55	66.27	32.1	30.02	13.38	0.32
	Median	63.73	66.24	64.82	64.04	32.16	30.47	1.91	0.19
	Std. Dev.	10.6	11.26	9.66	10.64	2.64	1.51	50.26	0.38
MS	Mean	69.5	69.06	68.44	68.83	31.14	29.97	6.21	0.8
	Median	68.81	69	68.6	68.41	31.58	29.87	3.9	0.48
	Std. Dev.	10.1	10.87	7.62	9.93	2.04	1.88	10.51	1.14
RW	Mean	85.55	87.04	72.38	94.86	30.76	29.63	14.99	0.74
	Median	89.26	88.08	73.56	93.45	30.98	29.75	3.35	0.43
	Std. Dev.	10.48	11.41	7.62	19.17	1.85	1.77	49.94	1.05
MW	Mean	88.93	89.3	72.82	98.61	30.63	29.34	15.06	0.79
	Median	89.13	88.8	73.48	95.25	30.97	29.12	3.44	0.39
	Std. Dev.	10.15	13.38	8.75	18.97	1.85	1.55	49.92	1.13

Table 1: This table presents descriptive values of the physiological measures over all participants and grouped per each condition (relaxed stationary - RS, MAT stationary - MS, relaxed walking - RW, MAT walking - MW) and device.

21 remaining participants was 28.9 ($SD = 4.5$) years; among them were 8 females and 13 males. During the recruitment process, it was ensured that participants were not diagnosed with any heart conditions, mental illnesses or learning disabilities. Likewise, all participants assured that they did not suffer from alcohol and/or drug addiction. Furthermore, they were asked to refrain from caffeine three hours before the experiment started. Participants were given a £15 gift voucher for taking part in the 1.5 hour long experiment session.

Initially, participants were introduced to the experiment environment at Body-Centric Lab of the Queen Mary University London. Before signing the consent form, they were briefed on the study background as well as the sensor placement on the body. Subsequently, the were asked to fill in an initial assessment consisting of demographic questions, self-reported fitness assessment, and smoking behavior as inquired in Weitkunat et al. [75]. Participants were given a short treadmill introduction and the mental arithmetic task was explained.

Next, the participants were asked to put on the chest-worn ECG sensors (Nexus 10 ECG with pre-gelled, disposable electrodes and Polar H7 chest belt). To ensure proper sensor fit, they were provided with visual material from the manufacturers on the correct sensor placement. The wrist-worn devices (Apple Watch and Microsoft Band 2), as well as finger skin conductance and skin temperature sensors were placed on the participants' left arm by the researcher. Correct data transmission for all sensors was initially checked by the researcher before the study started. During the experiment each participant was video recorded for traceability purposes given the participant's consent. Starting with the baseline condition, all participants were asked to remain seated for five minutes listening to meditation music via wireless headphones.

The conditions were assigned to each participant in counterbalanced order, alternating between *walking* and *stationary* while mental arithmetic tasks should be performed. This design has been followed for the last trial, while in the third trial participants were asked to walk while listening to relaxing music via wireless headphones. Please also refer to Figure 1 for a sketch of the study design depicting the sequence of conditions. Each trial (including the baseline) was followed by assessment of the SAM questionnaire including single-items on wake/tense arousal and perceived stressfulness of the task.

This study was reviewed and approved by the Ethical Committee of our institute.

RESULTS

Analyzing the physiological and subjective data from our participants, we will present the results of our statistical analysis following the structure of our hypothesized outcomes.

Data Preprocessing

For the analysis, we took a period of 4 minutes per each condition, meaning we excluded the first 50 and last 10 seconds due to novelty effects. Furthermore, we converted the Microsoft Band's provided skin resistance (R) measures ($kohms$) to match the unit of skin conductivity (G) provided by the Nexus device ($micro-mho$). We applied the following formula: $G = \frac{1}{R} * 1000$.

The descriptive measures (Mean, Median, Standard Deviation) for the recorded physiological data by each device and among all four conditions are presented in Table 1.

As proof of concept that the chosen study setup and task was stress-inducing, we compared the subjective stress and arousal in the different conditions; an overview is depicted in Figure 2. A comparison of the medians highlights that participants experienced higher arousal and stress in the MAT tasks compared to the relaxing-music tasks. These results indicate that the chosen tasks (MAT - mental arithmetic tasks) induced stress and, thus, we can expect to see a stress reaction in the devices' physiological data in the MS and MW conditions.

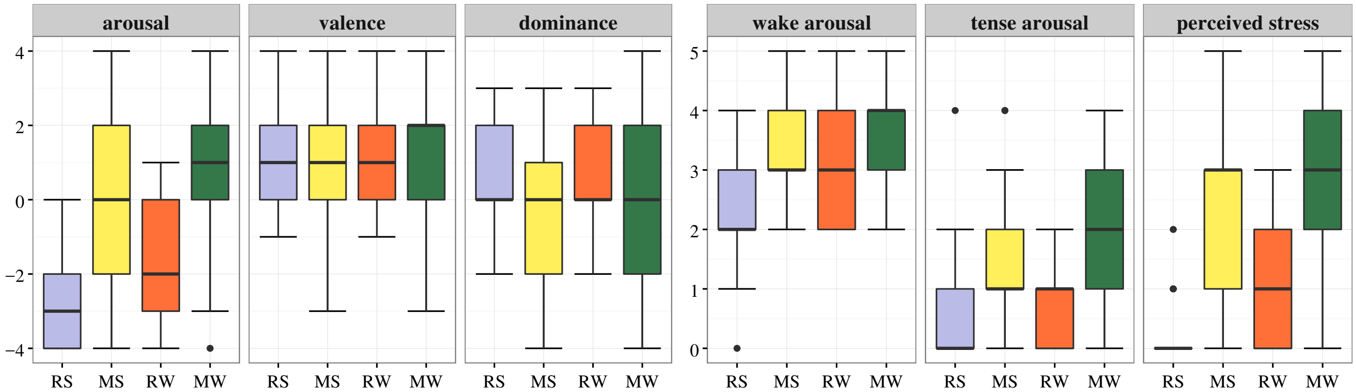

Figure 2: The Boxplot depicts the median values and inter-quartile range of the subjective measures for all participants and grouped among conditions (relaxed stationary - RS, MAT stationary - MS, relaxed walking - RW, MAT walking - MW). It suggests that the MAT conditions MS and MW were perceived more stressful.

H1: Investigating Physiological Data among Devices

We hypothesized differences in the physiological data under physical but not under stationary activity, hence, we investigated correlation referring to data accuracy and additionally performed Friedman and Wilcoxon Signed-Rank Tests for heart rate, skin temperature, and EDA for each physical activity condition.

After checking for normal-distribution of the physiological data, we performed Spearman correlations[7]. The results reveal moderate to strong correlations between the heart rate measures of the different devices over the whole data set. Considering Spearman's Rho for each physical condition separately, the physical activity showed to have a strong impact on the significance and strength of the device data correlating with each other.

Whereas in the stationary conditions, all devices correlated very strongly ($r_s > .95$, $p < 0.01$) with each other, there was only one strong correlation between the Polar and Nexus device and one moderate correlation between the Apple Watch and Polar under walking conditions. An overview of the correlation coefficients can be found in Table 2. For skin temperature measures, there was a moderate overall correlation between the Nexus and Microsoft Band ($r_s = .553, p = 0.000$). The correlation between the two devices was moderate in the stationary condition ($r_s = .537, p = 0.004$) and in the walking condition ($r_s = .472, p = 0.002$). For the electrodermal activity measures, we found a weak correlation between the Nexus and Microsoft Band ($r_s = .234, p = .037$). There were no other significant correlations found for the separate consideration of walking and stationary conditions.

		Apple Watch	Polar	Microsoft Band
Nexus	Overall	$.795^{**}$	$.889^{**}$	$.578^{**}$
	Stationary	$.989^{**}$	$.986^{**}$	$.966^{**}$
	Walking	NS	$.617^{**}$	NS
Apple Watch	Overall		$.851^{**}$	$.592^{**}$
	Stationary		$.993^{**}$	$.972^{**}$
	Walking		$.411^{*}$	NS
Polar	Overall			$.626^{**}$
	Stationary			$.977^{**}$
	Walking			NS

$^{**}p < .01$, $^{*}p < .05$, NS - not significant

Table 2: Spearman's Rho for the heart rate values of the 4 devices Nexus, Polar, Apple Watch and Microsoft Band.

Differences in Heart Rate

Testing on differences between the four devices regarding heart rate recording in the stationary activity condition, the Friedman Test revealed that there was no significant difference for heart rate amongst the devices; $\chi^2 = 4.286$ ($p = .232$). In contrast, significant differences were found for the walking condition indicating that the devices reported disparate heart rate readings; $\chi^2 = 43.133$ ($p = .000$). The post-hoc Wilcoxon Signed Rank Test with a Holm-Bonferroni correction[8] for the six comparisons were performed. It indicated no significant differences in the heart rate measures between the pairings of Nexus, Apple Watch and Polar. On the contrary, the Microsoft Band (MSB) reported a significant lower heart rate compared to the Nexus (N), Polar (P) and Apple Watch (AW); $Z_{N,MSB} = -4.773, p = .000$; $Z_{P,MSB} = -4.583, p = .000$; $Z_{AW,MSB} = -4.156, p = .000$.

Differences in Skin Temperature

For skin temperature, performing the Wilcoxon Signed-Rank Test indicated a significant difference between the Nexus and Microsoft Band among both physical activity conditions alike. The Microsoft Band showed a lower skin temperature in general regarding stationary condition ($Z = -4.503, p = .000$)

[7] The strength of correlation was determined as follows: 0.8-1.0 = very strong, 0.6-0.79 = strong, 0.4-0.59 = moderate, 0.2-0.39 = weak, and <0.2 = very weak, after Evans [18]

[8] Holm's sequential Bonferroni correction of $\alpha = 0.05$ resulted in $\alpha/6 = 0.0083, \alpha/5 = 0.01, \alpha/4 = .0125, \alpha/3 = .017, \alpha/2 = .025, \alpha/1 = .05$

		Polar	Apple Watch	Microsoft Band
Overall	Mean Error	6.84	8.28	12.06
	Std.	12.34	15.52	12.04
Stationary	Mean Error	**3.22**	3.42	5.44
	Std.	4.07	4.12	5.96
Walking	Mean Error	10.28	14.41	**19.03**
	Std.	16.03	21.37	12.87

Table 3: Average Error percentage of the heart rate signals compared to the Nexus 10 reference device. Highlighted are the lowest and highest error rate.

and walking condition ($Z = -4.256, p = .000$). On average, the Microsoft Band's reported skin temperature value was 1.31°C (Mdn = 1.40°C; $\sigma = 2.32$°C) lower than the Nexus skin temperature over all conditions.

Differences in Electrodermal Activity

Lastly, the Wilcoxon Signed-Rank Test revealed a significant difference of EDA measures between the Nexus and Microsoft Band among both physical activity conditions. The Microsoft Band showed a lower skin conductance in general with respect to the stationary condition - $Z = -5.125, p = .000$ and walking condition - $Z = -5.024, p = .000$. Here again, the Microsoft Band's reported EDA was 11.817 Micro-Mho (Mdn = 2.500 Micro-Mho; $\sigma = 44.188$ Micro-Mho) lower on average than the Nexus EDA over all conditions.

Error Rate of Heart Rate

Comparing the error rates to the laboratory measurement instrument (Nexus 10), revealed further differences among the two physical activity conditions. The error rate for every five second data window was calculated for each device d as

$$error_d = \frac{|hr_{Nexus} - hr_d|}{hr_{Nexus}} * 100$$

Considering the average error rates, the Polar chest belt performed best followed by the Apple Watch. In favor of our hypothesis, the error was higher in the walking conditions. The mean and standard deviation of those errors are presented in Table 3.

H2: Comparing Stress in Physical and Stationary Activity

According to our second hypothesis, we compared the physiological data for both, stressful and relaxed, situations under stationary activity and under physical activity. To test this, we performed two planned Wilcoxon Signed-Rank Tests for the relaxed and MAT conditions under the same physical activity. For the two planned comparisons, we applied a Bonferroni correction on $\alpha = 0.05$ which resulted in $\alpha/2 = .025$.

The Nexus reference device was able to pick up a significant increase in heart rate in the MAT condition while participants were seated; $Z = -2.381$, $p = 0.017$. None of the other heart rate monitors registered this change. Both the Nexus and Microsoft Band revealed differences in electrodermal activity while participants were stationary; $Z = -3.285$, $p = 0.001$ and

		arousal	wake arousal	tense arousal	perceiv. stress
Heart Rate	AW	.265*	.244*	NS	.252*
	MSB	.244*	NS	NS	NS
	Polar	.235*	.236*	NS	.248*
	Nexus	**.323****	**.277***	**NS**	**.284****
EDA	MSB	.361**	.272*	.337**	.376**
	Nexus	**.297****	**.362****	**NS**	**.296****
Skin Temp	MSB	NS	-.221*	NS	NS
	Nexus	**-.259***	**-.367****	**NS**	**-.262***

$^{**}p < .01, ^{*}p < .05$, NS - not significant

Table 4: Correlation of subjective measures and within-subject normalised physiological data from Nexus (highlighted), Polar, Microsoft Band (MSB) and Apple Watch (AW)

$Z = -3.058$, $p = 0.002$. Again in the stationary condition, a change in skin temperature was solely registered by the Nexus device; $Z = -2.416$, $p = 0.016$.

While walking on the treadmill performing MAT simultaneously, none of the devices was able to detect any changes in physiological data compared to the task where participants walking and listening to relaxing music.

H3: Correlating Subjective and Physiological Data

We hypothesized correlations between physiological data and subjectively perceived stress, arousal and valence. We performed Spearman correlations on the non-normally distributed data. To control for individual differences in the participants' heart rate, skin temperature and EDA responses, the physiological data was transformed using within-subject z-score standardization, as suggested by [7]. The results of the correlation are presented in Table 4.

The heart rate provided by Nexus, Apple Watch and Polar showed correlations with perceived arousal, wake arousal and stress. The Microsoft Band's heart rate showed mere correlation with the arousal measure. On the contrary, while EDA measures of both Nexus and Microsoft Band showed a weak agreement with arousal, wake arousal and stress, the Microsoft Band's EDA measure, additionally and as the only sensor source, correlated with tense arousal. The reference device's skin temperature measure correlated negatively with the arousal, wake arousal and tense arousal, while the Microsoft Band showed mere correlations of skin temperature with wake arousal.

Contrary to our hypothesis, none of the physiological data sources showed correlations with valence. Additionally, we were able to support our last hypothesis on the absence of a relationship between dominance and physiological data by applying the aforementioned statistical operations.

DISCUSSION

In this section, we discuss the results of our study with hindsight on our three hypotheses. Moreover, we present a *Design*

Space for using wearable devices in research settings, and further we conclude limitations of our study. We were able to proof our concept and study apparatus of inducing subjective stress with our implementation of the MAT task. The arousal, wake arousal and perceived stress ratings were significantly higher for the MAT tasks compared to the conditions where participants were listening to relaxing music.

Reliability of Devices in Different Physical Activity

According to Hypothesis 1a and 1b, we tested the differences amongst the sensing technologies and devices in both physical activities. We hypothesized that there would be no difference between the device data in stationary conditions but in walking conditions, due to decreased accuracy of wearables in movement. We could partly, and for a subset of devices and sensor streams, confirm both hypotheses.

Heart Rate

In the stationary conditions, we found strong correlations regarding heart rate values among all four devices. Furthermore, the Friedman Test showed no significant difference in the heart rate measures. This supports our Hypothesis 1a that heart rate values are consistent among the devices and highlights the accuracy of the devices in a stationary setting.

Contrary and supporting H 1b, our results show discrepancies in heart rate values recorded by different devices in the walking condition. All comparisons involving the Microsoft Band indicated significant differences in the recorded heart rate values, thus, we concluded this wrist-worn PPG heart rate sensor as the least accurate under movement. Under the walking conditions, a look at the average reported values of the Microsoft Band - as depicted in Table 1 - indicate that it tends to underreport the heart rate compared to the gold standard Nexus; on the contrary, the Apple Watch tends to reports higher heart rate values, though it was not significantly different. Further and in walking, both wrist-worn devices showed no correlations with the laboratory measurement instrument (Nexus 10) confirming that the PPG technology performs weaker under movement. As expected, the Polar H7 ECG chest strap performed closest to the Nexus ECG.

Skin Temperature

The Microsoft Band's reported skin temperature tended to be lower than the Nexus skin temperature by 1.31°C (±2.32°C) over all conditions and performed, hence and on first sight, against our hypotheses H1b. Considering the absolute skin temperature values, both devices (Microsoft Band and Nexus) provided inconsistent data through both physical activity conditions. On the contrary, the correlations between both devices were consistently strong, no matter of the physical activity. Both findings indicate, that these deviations in absolute values can be explained by the different sensor placements rather than an influence of physical activity; the Nexus sensor was placed on the upper forearm while the Microsoft Band was attached to the wrist.

Electrodermal Activity

The electrodermal activity data from the Microsoft Band and Nexus showed mere weak correlations over all conditions. But looking at the distinct physical activity conditions, there were no significant correlations which supports our first hypothesis and neglects the second. Further, there was a remarkable difference of 11.817 Micro-Mho (±44.188 Micro-Mho) over all conditions, regardless physical activity. The big variation can be partly explained with the sensitive skin conductance sensors loosing skin contact for small periods of time.

Error Rate of Heart Rate

We observed that error rate increased more than threefold for all devices in the walking conditions. This effect can be explained through an increased sensor noise in movement and an increased inaccuracy of PPG wrist-worn devices in higher heart rate ranges. Overall, the Polar ECG chest belt provided more accurate data compared to the wrist-worn PPG sensors. From the wrist-worn devices, the Apple Watch performed best in our study. This goes conform with findings from related work [63, 16]. Hence, this contributes to confirming our Hypothesis 1a; there is a difference in physiological data measured by different devices under movement.

Stress Related Changes in Physiological Measures

Following our Hypotheses 2a and H 2b, we investigated differences in physiological data indicating stress between between stationary and physical activity. The Nexus was the only device to fully support our Hypotheses 2a reporting significant differences in heart rate, electrodermal activity and skin temperature under stationary activity enabling us to trace a stress reactions. It detected a significant increase in heart rate and electrodermal activity during the MAT condition compared to the relaxed condition under stationary activity. Similar effects were observed for skin temperature measure. It showed the expected (according to [71]) decrease in skin temperature while performing the MAT compared to the relaxed condition in stationary activity. From the consumer range, the Microsoft Bands EDA sensor was the only one to show an increase in skin conductance while participants remained stationary. Neither the PPG wrist devices nor the Polar ECG chest belt were able to detect changes in heart rate data indicating stress responses. Due to the lack of accuracy of data recording provided by the tested devices under movement conditions, as discussed within Hypotheses 1a and 1b, we cannot fully test our Hypothesis 2b under movement.

Subjective Measures Linked to Physiological Data

Lastly, we hypothesized correlations between physiological measures recorded with our devices and the subjectively assessed measures (e.g. arousal, perceived stress). We hereby made the assumption that the physiological measures will correlate with valence, arousal, and perceived stress (H 3a-c). We further suggested, based on related literature, that there will be no correlations between dominance and the physiological sensor data (H 3d). Our results show that we are able to support all hypotheses, except H 3c - the correlation with valence. Hereby, all physiological measures recorded with the Nexus device revealed the strongest correlations what further strengthens the reliability of the Nexus kit as a suitable measurement tool. The Apple Watch and Polar heart rate sensor data hinted correlations regarding arousal, wake arousal and perceived stress, too. Microsoft Band's EDA sensor showed significant evidence for a correlation between all three arousal measures,

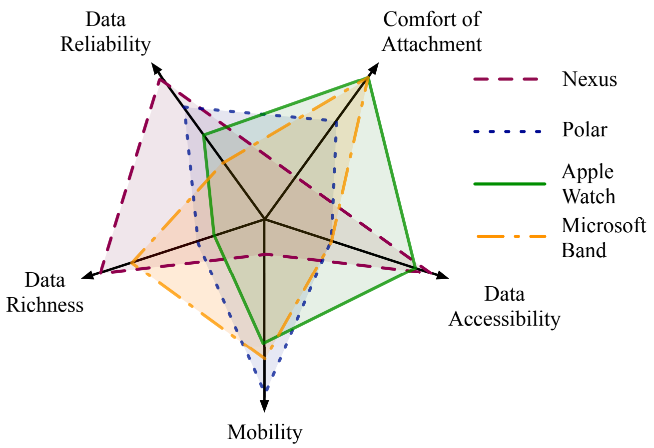

Figure 3: Illustrative schematic of the design space evaluation for our 4 test devices (Nexus, Polar, Apple Watch, Microsoft Band) in 5 criteria dimensions (data reliability, comfort of attachment, mobility, data richness, and data accessibility)

perceived stress and physiological data. On the contrary, its heart rate and skin temperature sensors did not perform well showing mere correlations regarding only one subjective measure. This is consistent with our previous results rendering the Microsoft Band the least reliable device in our study setup.

Design Space for Wearables Used in Research Settings

Based on our observations from the study and common evaluation criteria for wearable technology, like 'comfort' [33] or 'data reliability' [19], we inferred a *design space for wearables used in research settings* providing recommendations for suitable devices in different research scenarios. For this, we derived the following five dimensions partly grounded in the taxonomy by Khusainov et al. [36]: data reliability, comfort of attachment, mobility, data richness, and data accessibility. Lastly, we discuss the setting appropriateness of our wearables based on their specific advantages and disadvantages. An illustration of our assessment of our test devices can be found in Figure 3.

Data Reliability

Most of all, our and previous studies confirmed that there are variations in sensor data accuracy which results into a limited reliability. The Nexus device, as a laboratory tool, was the only device to show stress-related, statistically valid changes in heart rate, skin temperature and EDA. Wrist-worn, PPG-based devices tend to be less reliable in measuring heart rate than devices deriving heart rate values from ECG data. This effect is worsened in conditions involving physical movement. But there are even differences amongst devices using PPG technology. In our study, we identified the Microsoft Band 2 to be the most unreliable in terms of heart rate and skin temperature data while the Apple Watch performed acceptable. Surprisingly, the Microsoft Bands EDA sensor showed correlations with all subjective stress measures, which makes it a promising device for detecting stress. On the contrary, while heart rate chest belts with ECG technology proved to be more reliable than PPG sensors (i.e. [25]), we could not find significant differences in sensing data between stressed and relaxed conditions with the Polar device.

Comfort of Attachment

Comfort or wearability of wearables are not just an important factors for acceptance of the device [9], but play an important role for the study device choice. While the wrist wearables are designed to be worn all day long and are suitable for long-term in-situ studies due to the placement natural locations to wear technology [50], the Nexus and Polar are more purpose-led in their functionality and are designed to be worn for certain occasions. The Polar device is suitable for e.g. field studies due to its easy and quick attachment, but it can be visible through tight-fit clothing and may not be comfortable, especially for female participants, due to its placement. The Nexus, as a laboratory measurement tool with several applications, is relatively heavy (500 grams[9]) and requires detailed instructions on the correct placement of sensors. Therefore, it is cumbersome research settings requiring flexibility. Further, the self-adhesive stick on electrodes can cause discomfort when removed and may leave behind residue.

Mobility

A huge benefit of most wrist-worn devices is their mobility aspect. Without the need of cables, they allow the unconstrained movement of the participant. Additionally, their relatively long battery lifetime allows for them to be worn for a long time without the need to charge. While the Apple Watch promises a an 'all-day' battery life of 18 hours and the Microsoft Band 48 hours [31, 3], the Polar provides 400 hours of heart rate recording [17]. The Nexus promises more than 24 hours of operation [44]. All of the devices are advertised as wearable, but the Nexus would hardly be suitable for e.g. sleep studies, due to its bulky nature.

Data Richness

All of our test devices provided a different set of data varying in granularity. Looking at the heart rate measures alone, the Nexus provided a raw-ECG signal with a frequency of 256 Hz, while the Polar ECG chest belt did not allow access to the raw signal. On the contrary, the Apple Watch provided roughly one heart rate sample per second. Not just the granularity of a device is important, but also the diversity of sensors. The Microsoft Band is particularly richly equipped for a consumer device with e.g. heart rate, skin temperature, EDA, and UV sensors compared to other wrist-worn wearables.

Data Accessibility

Not just the richness of sensors is important, but also the ease of access to the data. The BioTrace+ software suite, which accompanies the Nexus, provides easy export and even real-time data visualizations making an access easy. Apple included HealthKit in their iOS system which allows CSV export of the collected heart rate samples. The Polar and Microsoft Bands sensor data is mere accessible through mobile APIs, which have to be included in a data collection app, or third-party applications. Here it becomes obvious that the ease of data accessibility needs to be improved.

[9] approximated weight by the manufacturer: www.mindmedia.info/CMS2014/products/systems/nexus-10-mkii

Advantages and Disadvantages
Considering the four named dimensions of the discussed devices, we illustratively summarized the fulfillment of each criteria per each device in Figure 3.

As can be seen, the Nexus kit covers three of the five dimensions and only lacks the *comfort of attachment* and *mobility* due to its bulkiness and the self-adhesive electrodes. If high data accuracy and richness is a prerequisite and the laboratory setting does not require much movement and physical activity from the participants, the Nexus kit serves as a reliable measurement tool for physiological data. It could hereby be suitable for stationary HCI studies, like e.g. desktop usability evaluations. On the contrary, *comfort of attachment* is an important criteria that needs to be considered; the more so when conducting studies with special groups e.g. children or mentally disabled people. Requiring wearables for non-stationary settings and field studies, e.g. for the evaluation of ambient interfaces, surely the PPG wrist devices provide the highest *comfort of attachment* and *mobility*.

A distinct disadvantage of ECG-based devices over wrist-PPG technology, is their data reliability. While the Microsoft Band proved to be the least reliable wrist-device in terms of heart rate and skin temperature, it showed to be rich in the provided sensor data and provides three relevant sensor for measuring stress responses and further studies on the reliability of a sensor fusion of this data are outstanding. The Apple Watch, which also lacks *data reliability*, though to a lesser degree than the Microsoft Band, provides better accessible data.

In terms of *data accessibility* the Polar chest belt performs poorly compared to Nexus and Apple Watch. Another drawback lies in *data richness* since it only assesses heart rate. Nevertheless, the Polar ECG chest belt serves as convenient alternative to the usually used laboratory devices. Its sufficient *data reliability*, easy attachment and *mobility* due to long battery lifetime make it suitable for long-term field studies, e.g. long-term effects of technology usage on stress.

Concluding, researchers should weigh the pros and cons for utilizing the discussed sensing technologies considering study setup, flexibility needed and purpose of the study.

Limitations
Although our results are giving important insights into the reliability of physiological data accessed by wearables, we tested only a limited amount of devices. Facing the variety of wearable (fitness) devices, our results may not apply for each of them and therefore are not generalizable. Further, the reliability of wrist-worn PPG heart rate sensors is influenced by factors, like skin pigmentation [65], which have not been assessed during the study. Our results are based on short-term data acquisition of approx. 20 minutes. It would be definitely interesting to validate the device performance in a longitudinal setting also including more participants. Since all participants were students with engineering background, there are implications on the performance during the mental arithmetic tasks (MAT). Although we could show by the subjectively assessed measures that participants felt more stressed in the MAT conditions, we did not track task performance i.e. error rate. A further investigation of participants' task performance and the adaptive adjustment of the MAT's difficulty would be interesting to observe also with respect to subjective and physiological stress measures.

CONCLUSION AND FUTURE WORK
By this work, we first contribute a comparison between PPG, wrist devices (Apple Watch, Microsoft Band 2) against an ECG chest strap (Polar H7 chest belt) and a laboratory measurement instrument with stick-on ECG technology (Nexus 10 kit) under different physical and stressful conditions. To evaluate the reliability of the named sensing technologies, we investigated the differences in physiological data measured by the devices (Hypotheses 1a and H1b) confirming that PPG-wearables tend to be less accurate in movement and the data gets less suitable for sensitive research settings. We further checked the influence of stress on physiological data under stationary and physical activity (Hypothesis 2) which could be only partly confirmed owed to the lack of accuracy in the devices. As another contribution, we could show that perceived stress and arousal (tense and wake) correlate with the physiological data suggesting a strong relation between physiological and subjectively felt stress, whereas there no correlations for valence and dominance observed (Hypotheses 3). Based on our findings, we lastly contribute a *Design Space for Wearables Used in Research Settings* addressing five dimension covering important criteria for choosing an appropriate measurement tool for research purposes.

In future work, we plan to investigate noise reduction by using the accelerometer data, which is readily available in most consumer devices. Therefore, we will compare more wearables involving new products using improved sensors and data extraction algorithms. Also most of these wearables are not scientifically validated for their accuracy and validity. Novel consumer devices even target well-being aspects and stress such as the Garmin Vivosmart 3[10], which claims to use HRV to calculate a proprietary stress score throughout the day. In terms of stress and emotion detection, we plan to have a closer look at stress detection through wearables in the wild as there are already approaches based on mobile sensing data [10, 40]. The combination of those approaches with wearable physiological data could lead to more accurate predictions and models [20].

By this work we believe to have presented a first step towards assessing sensing technologies in wearables for their reliability and accuracy, as well as having provided fruitful insights for other researchers when it comes to decide which measurement tool to use in a study.

ACKNOWLEDGMENTS
This work was kindly supported by the Centre of Intelligent sensing of the Queen Mary University London. Hamed Haddadi was partially funded by EPSRC Databox grant (Ref: EP/N028260/1) and EPSRC IoT-in-the-Wild grant (Ref: EP/L023504/1). This work was partly conducted within the Amplify project funded from the European Research Council (ERC) (grant agreement no. 683008).

[10] www.garmin.com

REFERENCES

1. John Allen. 2007. Photoplethysmography and Its Application in Clinical Physiological Measurement. *Physiological Measurement* 28, 3 (Mar 2007), R1–R39. DOI:http://dx.doi.org/10.1088/0967-3334/28/3/R01
2. Stein Andersson and Arnstein Finset. 1998. Heart rate and skin conductance reactivity to brief psychological stress in brain-injured patients. *Journal of Psychosomatic Research* 44, 6 (1998), 645 – 656. DOI: http://dx.doi.org/10.1016/S0022-3999(97)00305-X
3. Michael Andronico. 2017. Microsoft Band 2 vs. Apple Watch, Fitbit Surge and Garmin Vivoactive. Accessed: 15/12/2017. (2017). https://www.tomsguide.com/us/microsoft-band-vs-apple-watch,news-21684.html
4. Lawrence M Baker and William M Taylor. 1954. The relationship under stress between changes in skin temperature, electrical skin resistance, and pulse rate. *Journal of experimental psychology* 48, 5 (1954), 361.
5. Armando Barreto, Jing Zhai, and Malek Adjouadi. 2007. Non-intrusive physiological monitoring for automated stress detection in human-computer interaction. *Human–Computer Interaction* 4796 (2007), 29–38.
6. A. B. Barreto, S. D. Scargle, and M Adjouadi. 1999. A real-time assistive computer interface for users with motor disabilities. *ACM SIGCAPH Computers and the Physically Handicapped* 64 (1999), 6–16.
7. Gershon Ben-Shakhar. 1985. Standardization Within Individuals: a Simple Method to Neutralize Individual Differences in Skin Conductance. *Psychophysiology* 22, 3 (May 1985), 292–299. DOI: http://dx.doi.org/10.1111/j.1469-8986.1985.tb01603.x
8. Irving Biederman. 1973. Mental set and mental arithmetic. *Memory & Cognition* 1, 3 (1973), 383–386.
9. Kerry Bodine and Francine Gemperle. 2003. Effects of Functionality on Perceived Comfort of Wearables. *ISWC* (2003), 57–60. DOI: http://dx.doi.org/10.1109/ISWC.2003.1241394
10. Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. 2014. Daily Stress Recognition From Mobile Phone Data, Weather Conditions and Individual Traits. In *Proceedings of the ACM International Conference on Multimedia (MM'14)*. 477–486. DOI: http://dx.doi.org/10.1145/2647868.2654933
11. Margaret M Bradley and Peter J Lang. 1994. Measuring Emotion: the Self-Assessment Manikin and the Semantic Differential. *Journal Of Behavior Therapy And Experimental Psychiatry* 25, 1 (Mar 1994), 49–59.
12. Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A Global Measure of Perceived Stress. *Journal of Health and Social Behavior* 24, 4 (1983), 385–396. http://www.jstor.org/stable/2136404
13. Richard J. Davidson, Klaus R. Scherer, and H. Hill Goldsmith. 2003. *Handbook of Affective Sciences (pp. 572-595)*. New York: Oxford University Press.
14. Michael E Dawson, Anne M Schell, and Diane L Filion. 2007. The electrodermal system. *Handbook of psychophysiology* 2 (2007), 200–223.
15. M Eid, P Notz, P Schwenkmezger, and R Steyer. 1994. Sind Stimmungsdimensionen monopolar? Ein Überblick über empirische Befunde und Untersuchungen mit faktorenanalytischen Modellen für kontinuierliche und kategoriale Variablen sowie neuere Ergebnisse. *Zeitschrift für Differentielle und Diagnostische Psychologie* 15, 4 (1994), 211–233.
16. Fatema El-Amrawy and Mohamed Ismail Nounou. 2015. Are Currently Available Wearable Devices for Activity Tracking and Heart Rate Monitoring Accurate, Precise, and Medically Beneficial? *Healthcare Informatics Research* 21, 4 (2015), 315. DOI: http://dx.doi.org/10.4258/hir.2015.21.4.315
17. Polar Electro. 2017. How to check the battery level status of my heart rate sensor? Accessed: 15/12/2017. (2017). https://support.polar.com/en/support/how_to_check_the_battery_level_status_of_my_heart_rate_sensor
18. J. D. Evans. 1996. *Straightforward statistics for the behavioral sciences*. Brooks/Cole Publishing.
19. Kelly R Evenson, Michelle M Goto, and Robert D Furberg. 2015. Systematic Review of the Validity and Reliability of Consumer-Wearable Activity Trackers. *International Journal of Behavioral Nutrition and Physical Activity* 12, 1 (Dec 2015), e192. DOI: http://dx.doi.org/10.1186/s12966-015-0314-1
20. Anja Exler, Andrea Schankin, Christoph Klebsattel, and Michael Beigl. 2016. A Wearable System for Mood Assessment Considering Smartphone Features and Data From Mobile ECGs. In *Adjunct Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2016 ACM International Symposium on Wearable Computers (Ubicomp/ISWC'16 Adjunct)*. DOI: http://dx.doi.org/10.1145/2968219.2968302
21. Raihana Ferdous, Venet Osmani, and Oscar Mayora. 2015. Smartphone App Usage as a Predictor of Perceived Stress Levels at Workplace. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '15)*. 225–228. DOI:http://dx.doi.org/10.4108/icst.pervasivehealth.2015.260192
22. Aaron J Fisher and Michelle G Newman. 2013. Heart rate and autonomic response to stress after experimental induction of worry versus relaxation in healthy, high-worry, and generalized anxiety disorder individuals. *Biological psychology* 93, 1 (2013), 65–74.
23. Thomas B Fitzpatrick. 1988. The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Archives of Dermatology* 124, 6 (1988), 869. DOI: http://dx.doi.org/10.1001/archderm.1988.01670060015008

24. Peter J Gianaros, Ikechukwu C Onyewuenyi, Lei K Sheu, Israel C Christie, and Hugo D Critchley. 2012. Brain systems for baroreflex suppression during stress in humans. *Human brain mapping* 33, 7 (2012), 1700–1716.

25. Stephen Gillinov, Muhammad Etiwy, Robert Wang, Gordon Blackburn, Dermot Phelan, A Marc Gillinov, Penny Houghtaling, Hoda Javadikasgari, and Milind Y Desai. 2017. Variable Accuracy of Wearable Heart Rate Monitors During Aerobic Exercise. *Medicine & Science in Sports & Exercise* 49, 8 (Aug 2017), 1697–1703. DOI: http://dx.doi.org/10.1249/MSS.0000000000001284

26. Christian Hamilton-Craig, Allison Fifoot, Mark Hansen, Matthew Pincus, Jonathan Chan, Darren L. Walters, and Kelley R. Branch. 2014. Diagnostic performance and cost of CT angiography versus stress ECG – A randomized prospective study of suspected acute coronary syndrome chest pain in the emergency department (CT-COMPARE). *International Journal of Cardiology* 177, 3 (2014), 867 – 873. DOI: http://dx.doi.org/10.1016/j.ijcard.2014.10.090

27. Skjalg S Hassellund, Arnljot Flaa, Leiv Sandvik, Sverre E Kjeldsen, and Morten Rostrup. 2010. Long-Term Stability of Cardiovascular and Catecholamine Responses to Stress Tests An 18-Year Follow-Up Study. *Hypertension* 55, 1 (2010), 131–136.

28. Jennifer A Healey and Rosalind W Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems* 6, 2 (2005), 156–166.

29. Javier Hernandez, Rob R Morris, and Rosalind W Picard. 2011. Call center stress recognition with person-specific models. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 125–134.

30. Gabriella Ilie and William Forde Thompson. 2006. A Comparison of Acoustic Cues in Music and Speech for Three Dimensions of Affect. *Music Perception: An Interdisciplinary Journal* 23, 4 (Apr 2006), 319–330. DOI:http://dx.doi.org/10.1525/mp.2006.23.4.319

31. Apple Inc. 2017a. Apple Watch Series 3 Battery Information. Accessed: 15/12/2017. (2017). https://www.apple.com/uk/watch/battery.html

32. Statista Inc. 2017b. Forecasted value of the global wearable devices market from 2012 to 2018. Accessed: 16/08/2017. (2017). https://www.statista.com/statistics/302482/wearable-device-market-value/

33. Martin Jagelka, Martin Donoval, Peter Telek, František Horìnek, Martin Weis, and Martin Dařiček. 2016. Wearable Healthcare Electronics for 24-7 Monitoring with Focus on User Comfort. In *2016 26th International Conference Radioelektronika (RADIOELEKTRONIKA*. 5–9. DOI: http://dx.doi.org/10.1109/RADIOELEK.2016.7477444

34. Edward Jo, Kiana Lewis, Dean Directo, Michael J Kim, and Brett A Dolezal. 2016. Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking. *Journal of Sports Science & Medicine* 15, 3 (Sep 2016), 540.

35. Hisanori Kataoka, Hiroshi Kano, Hiroaki Yoshida, Atsuo Saijo, Masashi Yasuda, and Masato Osumi. 1998. Development of a skin temperature measuring system for non-contact stress evaluation. In *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, Vol. 2. IEEE, 940–943.

36. Rinat Khusainov, Djamel Azzi, Ifeyinwa E Achumba, and Sebastian D Bersch. 2013. Real-time human ambulation, activity, and physiological monitoring: Taxonomy of issues, techniques, applications, challenges and limitations. *Sensors* 13, 10 (2013), 12852–12902.

37. Arthur F Kramer. 1991. Physiological metrics of mental workload: A review of recent progress. *Multiple-task performance* (1991), 279–328.

38. Jan Kučera, James Scott, and Nicholas Chen. 2017. Probing Calmness in Applications Using a Calm Display Prototype. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp '17)*. ACM, New York, NY, USA, 965–969. DOI:http://dx.doi.org/10.1145/3123024.3124564

39. Wenhui Liao, Weihong Zhang, Zhiwei Zhu, and Qiang Ji. 2005. A real-time human stress monitoring system using dynamic Bayesian network. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*. IEEE, 70–70.

40. Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. MoodScope: Building a Mood Sensor From Smartphone Usage Patterns. In *Proceeding of the 11th annual international conference (MobiSys '13)*. 389–402. DOI:http://dx.doi.org/10.1145/2462456.2464449

41. Wan-Hua Lin, Dan Wu, Chunyue Li, Heye Zhang, and Yuan-Ting Zhang. 2014. *Comparison of Heart Rate Variability From PPG with That From ECGIFMBE Proceedings*. 213–215. DOI: http://dx.doi.org/10.1007/978-3-319-03005-0_54

42. Wolfgang Linden. 1991. What do arithmetic stress tests measure? Protocol variations and cardiovascular responses. *Psychophysiology* 28, 1 (1991), 91–102.

43. Daniel J McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W Picard. 2016. Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4000–4004.

44. Mind Media. 2017. Nexus-10 MKII Specifications. Accessed: 15/12/2017. (2017). https://www.mindmedia.com/products/nexus-10-mkii/

45. Jon D Morris. 1995. Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research* 35, 6 (1995), 63–68.

46. Mats Najström and Billy Jansson. 2007. Skin conductance responses as predictor of emotional responses to stressful life events. *Behaviour Research and Therapy* 45, 10 (2007), 2456 – 2463. DOI: http://dx.doi.org/10.1016/j.brat.2007.03.001

47. Jakub Parak and Ilkka Korhonen. 2014. Evaluation of Wearable Consumer Heart Rate Monitors Based on Photopletysmography. In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 3670–3673. DOI: http://dx.doi.org/10.1109/EMBC.2014.6944419

48. Rosalind W Picard and Jennifer Healey. 1997. Affective wearables. In *Wearable Computers, 1997. Digest of Papers., First International Symposium on*. IEEE, 90–97.

49. Julia Pietilä, Saeed Mehrang, Johanna Tolonen, Elina Helander, Holly Jimison, Misha Pavel, and Ilkka Korhonen. 2017. *Evaluation of the Accuracy and Reliability for Photoplethysmography Based Heart Rate and Beat-to-Beat Detection During Daily ActivitiesIFMBE Proceedings*. 145–148. DOI: http://dx.doi.org/10.1007/978-981-10-5122-7_37

50. Halley Profita, James Clawson, Scott M Gilliland, Clint Zeagler, Thad Starner, Jim Budd, and Ellen Yi-Luen Do. 2013. Don't Mind Me Touching My Wrist: a Case Study of Interacting with on-Body Technology in Public. *ISWC* (2013), 89. DOI: http://dx.doi.org/10.1145/2493988.2494331

51. Andrea A Quesada, Rosana M Tristao, Riccardo Pratesi, and Oliver T Wolf. 2014. Hyper-responsiveness to acute stress, emotional problems and poorer memory in former preterm children. *Stress* 17, 5 (2014), 389–399.

52. Michaela Riediger, Cornelia Wrzus, Kathrin Klipker, Viktor Muller, FLorian Schmiedek, and Gert G Wagner. 2014. Outside of the Laboratory: Associations of Working-Memory Performance with Psychological and Physiological Arousal Vary with Age. *Psychology and Aging* 29, 1 (2014), 103–114. DOI: http://dx.doi.org/10.1037/a0035766

53. James A Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

54. Valorie N Salimpoor, Mitchel Benovoy, Gregory Longo, Jeremy R Cooperstock, and Robert J Zatorre. 2009. The rewarding aspects of music listening are related to degree of emotional arousal. *PloS one* 4, 10 (2009), e7487.

55. Hillary S Schaefer, Christine L Larson, Richard J Davidson, and James A Coan. 2014. Brain, body, and cognition: Neural, physiological and self-report correlates of phobic and normative fear. *Biological psychology* 98 (2014), 59–69.

56. Axel Schäfer and Jan Vagedes. 2013. How Accurate Is Pulse Rate Variability as an Estimate of Heart Rate Variability? *International Journal of Cardiology* 166, 1 (Jun 2013), 15–29. DOI: http://dx.doi.org/10.1016/j.ijcard.2012.03.119

57. Jocelyn Scheirer, Raul Fernandez, Jonathan Klein, and Rosalind W Picard. 2002. Frustrating the user on purpose: a step toward building an affective computer. *Interacting with computers* 14, 2 (2002), 93–118.

58. Ulrich Schimmack and Alexander Grob. 2000. Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality* 14, 4 (2000), 325–345.

59. Neil Schneiderman, Gail Ironson, and Scott D Siegel. 2005. Stress and health: psychological, behavioral, and biological determinants. *Annu. Rev. Clin. Psychol.* 1 (2005), 607–628.

60. N Selvaraj, A Jaryal, J Santhosh, K K Deepak, and S Anand. 2008. Assessment of Heart Rate Variability Derived From Finger-Tip Photoplethysmography as Compared to Electrocardiography. *Journal of Medical Engineering & Technology* 32, 6 (Jul 2008), 479–484. DOI:http://dx.doi.org/10.1080/03091900701781317

61. Peter Seraganian, Attila Szabo, and Thomas G Brown. 1997. The effect of vocalization on the heart rate response to mental arithmetic. *Physiology & behavior* 62, 2 (1997), 221–224.

62. Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. 2010. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on information technology in biomedicine* 14, 2 (2010), 410–417.

63. A Shcherbina, C M Mattsson, and D Waggott. 2017. Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine* 7, 2 (2017), 3. DOI: http://dx.doi.org/10.3390/jpm7020003

64. RP Sloan, PA Shapiro, E Bagiella, SM Boni, M Paik, JT Bigger, RC Steinman, and JM Gorman. 1994. Effect of mental stress throughout the day on cardiac autonomic control. *Biological psychology* 37, 2 (1994), 89–99.

65. D K Spierer, Z Rosen, and L L Litman. 2015. Validation of Photoplethysmography as a Method to Detect Heart Rate During Rest and Exercise. *Journal of Medical Engineering & Technology* 39, 5 (2015), 264–271. DOI: http://dx.doi.org/10.3109/03091902.2015.1047536

66. Sarah E Stahl, Hyun-Sung An, Danae M Dinkel, John M Noble, and Jung-Min Lee. 2016. How Accurate Are the Wrist-Based Heart Rate Monitors During Walking and Running Activities? Are They Accurate Enough? *BMJ Open Sport & Exercise Medicine* 2, 1 (Apr 2016), e000106. DOI: http://dx.doi.org/10.1136/bmjsem-2015-000106

67. Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin L Griss. 2010. Activity-Aware Mental Stress Detection Using Physiological Sensors. *MobiCASE* 76, 2 (2010), 282–301. DOI:http://dx.doi.org/10.1007/978-3-642-29336-8_16

68. Toshiyo Tamura, Yuka Maeda, Masaki Sekine, and Masaki Yoshida. 2014. Wearable Photoplethysmographic Sensors – Past and Present. *Electronics* 3, 2 (Jun 2014), 282–302. DOI: http://dx.doi.org/10.3390/electronics3020282

69. Robert E Thayer. 1998. *The Biopsychology of Mood and Arousal.*

70. Joe Tomaka, Jim Blascovich, and Laura Swart. 1994. Effects of vocalization on cardiovascular and electrodermal responses during mental arithmetic. *International Journal of Psychophysiology* 18, 1 (1994), 23–33.

71. Christiaan H Vinkers, Renske Penning, Juliane Hellhammer, Joris C Verster, John HGM Klaessens, Berend Olivier, and Cor J Kalkman. 2013. The effect of stress on core and peripheral body temperature in humans. *Stress* 16, 5 (2013), 520–530.

72. Elke Vlemincx, Ilse Van Diest, and Omer Van den Bergh. 2012. A Sigh Following Sustained Attention and Mental Stress: Effects on Respiratory Variability. *Physiology & Behavior* 107, 1 (Aug 2012), 1–6. DOI: http://dx.doi.org/10.1016/j.physbeh.2012.05.013

73. Matthew P Wallen, Sjaan R Gomersall, Shelley E Keating, Ulrik Wisløff, and Jeff S Coombes. 2016. Accuracy of Heart Rate Watches: Implications for Weight Management. *PLoS ONE* 11, 5 (May 2016), e0154420. DOI:http://dx.doi.org/10.1371/journal.pone.0154420

74. Christoph Weinert. 2016. Coping with the Dark Side of IT Usage: Mitigating the Effect of Technostress. In *Proceedings of the 2016 ACM SIGMIS Conference on Computers and People Research (SIGMIS-CPR '16).* ACM, New York, NY, USA, 9–10. DOI: http://dx.doi.org/10.1145/2890602.2906189

75. Rolf Weitkunat, Christopher RE Coggins, Zheng Sponsiello-Wang, Gerd Kallischnigg, and Ruth Dempsey. 2013. Assessment of cigarette smoking in epidemiologic studies. *Beiträge zur Tabakforschung/Contributions to Tobacco Research* 25, 7 (2013), 638–648.

76. Charlotte VO Witvliet and Scott R Vrana. 2007. Play it again Sam: Repeated exposure to emotionally evocative music polarises liking and smiling responses, and influences other affective reports, facial EMG, and heart rate. *Cognition and Emotion* 21, 1 (2007), 3–25.