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 market1 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 tech-nostress [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 Jawbone2 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 3) against a heart rate chest strap (Polar H74) and a laboratory measurement instrument (Nexus 10 kit) under different physical and stressful conditions. To evaluate the

1The market value for wearable devices is forecasted to be almost 6 billion USD by 2018 [32].
2www.fitbit.com and www.jawbone.com
4www.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 wear-
abless for research settings with a focus on stress assessment.
We therefore provide background information on physiologi-
cal 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 phys-
iological 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 in-
creasingly popular and ubiquitous. While early devices fo-
cused merely on activity tracking via step count or flights of
stairs climbed, modern devices incorporated additional bio-
physiological 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 sys-
tems 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 de-
tecting 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. De-
pending 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 detection; 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 com-
pared 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 dimen-
sions: 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 in-
creased 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. 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 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.
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].

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 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 to 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 participants was 25, ranging from 21-27.

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.
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: $G = \frac{1}{R} \times 1000$.

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.

<table>
<thead>
<tr>
<th></th>
<th>Heart Rate</th>
<th>Skin Temperature</th>
<th>EDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nexus</td>
<td>Polar</td>
<td>Microsoft Band</td>
</tr>
<tr>
<td><strong>RS</strong></td>
<td>Mean 66.16</td>
<td>66.89</td>
<td>66.55</td>
</tr>
<tr>
<td></td>
<td>Median 63.73</td>
<td>66.24</td>
<td>64.82</td>
</tr>
<tr>
<td></td>
<td>Std. Dev. 10.6</td>
<td>11.26</td>
<td>9.66</td>
</tr>
<tr>
<td><strong>MS</strong></td>
<td>Mean 69.5</td>
<td>69.06</td>
<td>68.44</td>
</tr>
<tr>
<td></td>
<td>Median 68.81</td>
<td>69</td>
<td>68.6</td>
</tr>
<tr>
<td></td>
<td>Std. Dev. 10.1</td>
<td>10.87</td>
<td>7.62</td>
</tr>
<tr>
<td><strong>RW</strong></td>
<td>Mean 85.55</td>
<td>87.04</td>
<td>72.38</td>
</tr>
<tr>
<td></td>
<td>Median 89.26</td>
<td>88.08</td>
<td>73.56</td>
</tr>
<tr>
<td></td>
<td>Std. Dev. 10.48</td>
<td>11.41</td>
<td>7.62</td>
</tr>
<tr>
<td><strong>MW</strong></td>
<td>Mean 88.93</td>
<td>89.3</td>
<td>72.82</td>
</tr>
<tr>
<td></td>
<td>Median 89.13</td>
<td>88.8</td>
<td>73.48</td>
</tr>
<tr>
<td></td>
<td>Std. Dev. 10.15</td>
<td>13.38</td>
<td>8.75</td>
</tr>
</tbody>
</table>

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.
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 (\(r_s > .95, p < .01\)) 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 (\(r_s = .553, p = .000\)). The correlation between the two devices was moderate in the stationary condition (\(r_s = .537, p = .004\)) and in the walking condition (\(r_s = .472, p = .002\)). 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.

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; \(\chi^2 = 4.286, (p = .232)\).

\(^7\)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 \([18]\)

<table>
<thead>
<tr>
<th></th>
<th>Apple Watch</th>
<th>Polar</th>
<th>Microsoft Band</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nexus</strong></td>
<td>Overall</td>
<td>.795**</td>
<td>.889**</td>
</tr>
<tr>
<td></td>
<td>Stationary</td>
<td>.989**</td>
<td>.986**</td>
</tr>
<tr>
<td></td>
<td>Walking</td>
<td>NS</td>
<td>.617**</td>
</tr>
<tr>
<td><strong>Apple Watch</strong></td>
<td>Overall</td>
<td>.851**</td>
<td>.592**</td>
</tr>
<tr>
<td></td>
<td>Stationary</td>
<td>.993**</td>
<td>.972**</td>
</tr>
<tr>
<td></td>
<td>Walking</td>
<td>.411*</td>
<td>NS</td>
</tr>
<tr>
<td><strong>Polar</strong></td>
<td>Overall</td>
<td>NS</td>
<td>.626**</td>
</tr>
<tr>
<td></td>
<td>Stationary</td>
<td>.977**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Walking</td>
<td>NS</td>
<td></td>
</tr>
</tbody>
</table>

\(^* p < .01, ^{**} p < .05, NS - not significant\)

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

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); \(Z_{N,MSB} = -4.773, p = .000; Z_{PMSB} = -4.583, p = .000; Z_{AW,MSB} = -4.156, p = .000\).

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 (\(Z = -4.503, p = .000\))

\(^{8}\)Holm’s sequential Bonferroni correction of \(\alpha = 0.05\) resulted in \(\alpha/6 = 0.0083, \alpha/5 = 0.01, \alpha/4 = .0125, \alpha/3 = .017, \alpha/2 = .025, \alpha/1 = .05\)
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; $\sigma = 44.18$ 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

$$error_d = \frac{|hr_{d - Nexus} - hr_{d}|}{hr_{d - Nexus}} \times 100$$

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.

<table>
<thead>
<tr>
<th></th>
<th>Polar</th>
<th>Apple Watch</th>
<th>Microsoft Band</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Error</td>
<td>6.84</td>
<td>8.28</td>
<td>12.06</td>
</tr>
<tr>
<td>Std.</td>
<td>12.34</td>
<td>15.52</td>
<td>12.04</td>
</tr>
<tr>
<td><strong>Stationary</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Error</td>
<td>3.22</td>
<td>3.42</td>
<td>5.44</td>
</tr>
<tr>
<td>Std.</td>
<td>4.07</td>
<td>4.12</td>
<td>5.96</td>
</tr>
<tr>
<td><strong>Walking</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Error</td>
<td>10.28</td>
<td>14.41</td>
<td>19.03</td>
</tr>
<tr>
<td>Std.</td>
<td>16.03</td>
<td>21.37</td>
<td>12.87</td>
</tr>
</tbody>
</table>

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.

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 $Z = -4.812, p = .000$.

For the two planned comparisons, we applied a Bonferroni correction on $\alpha = 0.05$ which resulted in $\alpha/2 = .025$.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heart Rate</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polar</td>
<td>.235**</td>
<td>.236</td>
<td>.248**</td>
</tr>
<tr>
<td>Nexus</td>
<td>.323**</td>
<td>.277**</td>
<td>.376**</td>
</tr>
<tr>
<td><strong>EDA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polar</td>
<td>.361**</td>
<td>.272</td>
<td>.337**</td>
</tr>
<tr>
<td>Nexus</td>
<td>.297**</td>
<td>.362**</td>
<td>.296**</td>
</tr>
<tr>
<td><strong>Skin Temp</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nexus</td>
<td>-.259**</td>
<td>-.367**</td>
<td>-.262**</td>
</tr>
</tbody>
</table>

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

$** p < .01, * p < .05, NS - not significant$
were consistently strong, no matter of the physical activity. According to Hypothesis 1a and 1b, we tested the differences 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 under-report the heart rate compared to the gold standard Nexus; on the contrary, the Apple Watch tends to report 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 (±4.188 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 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 reaction. 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.
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 sensors 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 factor 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 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, 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


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](http://dx.doi.org/10.1007/978-3-319-03005-0_54)


