Comparing apples and oranges or different types of citrus fruits?

Using wearable vs. stationary devices to analyze psychophysiological data

Konstantinou Pinelopi, MSc[1]; Trigeorgi Andria, MSc[2]; Georgiou Chryssis, PhD[2] Gloster Andrew T., PhD[3], Panayiotou Georgia, PhD[1] & Karekla Maria, PhD[1]

[1]Department of Psychology, University of Cyprus, Nicosia, Cyprus

[2]Department of Computer Science, University of Cyprus, Nicosia, Cyprus

[3]Department of Psychology, University of Basel, Switzerland

Corresponding author: Maria Karekla, Ph.D., Department of Psychology, University of Cyprus, P.O. Box 20537, Nicosia 1678, Cyprus; TEL: 357 22 892100; mkarekla@ucy.ac.cy

Abstract

Wearable devices capable of capturing psychophysiological signals are popular. However, such devices have yet to be established in experimental and clinical research. This study, therefore, compared psychophysiological data (skin conductance level (SCL), heart rate (HR), and heart rate variability (HRV)) captured with a wearable device (Microsoft band 2) to those of a stationary device (Biopac MP150), in an experimental pain induction paradigm. Additionally, the present study aimed to compare two analytical techniques of HRV psychophysiological data: traditional (i.e., peaks are detected and manually checked) vs. automated analysis using Python programs. Forty-three university students (86% female; *Mage* = 21.37 years) participated in the cold-pressor pain induction task. Results showed that the majority of the correlations between the two devices for the mean HR were significant and strong (*rs* > .80) both during baseline and experimental phases. For the time-domain measure of mean RR (function of autonomic influences) of HRV, the correlations between the two devices at baseline were almost perfect (*rs* = .99) whereas at the experimental phase were significantly strong (*rs* > .74). However, no significant correlations were found for mean SCL (*p* > .05). Additionally, automated analysis led to similar features for HRV stationary data as the traditional analysis. Implications for data collection include the establishment of a methodology to compare stationary to mobile devices and a new, more cost efficient way of collecting psychophysiological data. Implications for data analysis include analyzing the data faster, with less effort and allowing for large amounts of data to be recorded.

Keywords: wearable device; psychophysiological data; stationary equipment; heart rate variability; skin conductance response; Bland-Altman plots; Root Mean Square Error

1. Introduction

Traditionally, measurement of psychophysiological data took place in the laboratory using stationary equipment. Measurement of psychophysiological data in this way contributed important findings in psychology, especially in the area of emotions research (Ries, Touryan, Vettel, McDowell, & Hairston, 2014). However, psychophysiological research had been confined to the study of signals only in controlled laboratory environments. Consequently, laboratory experiments have been criticized for their analogue nature and for placing participants in artificial settings in which all aspects are controlled by the researcher and are detached from the participants' natural environment and real life (Reis & Gosling, 2010). Findings do not represent situations, behaviors and actions from real life but instead may give a distorted picture of them (Vissers, Heyne, Peters, & Guerts, 2001). There is a plethora of information that can be learned if we explore psychophysiological signals within a persons' environment instead of examining them only in analogue experimental conditions (e.g., psychophysiological responses of a person during stress, pain, anger, etc.).

Recent technological advances led to the creation of new wearable devices. Wearable devices are technological devices which can be worn on the human body as accessories or incorporated into clothing, and are able to measure psychophysiological signals anywhere and anytime (Goncu-berk & Topcuoglu, 2017). Specifically, wearable devices can measure physiological signals such as heart rate (HR), heart rate variability (HRV), respiration rate, skin and body temperature, electrodermal activity (EDA), galvanic skin response (GSR), electromyography (EEG; Taj-Eldin, Ryan, O'flynn, & Galvin, 2018), etc. They have the advantages of being noninvasive, mobile, unobtrusive, less costly, easy to use, and aesthetically appealing to the individuals wearing them, compared to stationary devices. Wearable devices claim to collect the same or similar psychophysiological signals as stationary equipment but in an easier, faster and effortless way. For example, it requires more effort and time for researchers to place the electrodes on participants' arms, face and fingers, than to wear the wearable device. Importantly, they can measure psychophysiological signals continuously, while people go about their daily lives (Garbarino, Lai, Bender, Picard, & Tognetti, 2014; Ragot, Martin, Em, Pallamin, & Diverrez, 2017).

Despite the rise in popularity of such devices (e.g., for measuring HR during exercise, measuring steps taken, etc.), their use for scientific purposes and assessment of psychophysiological signals is still in its infancy. This is due in part to the lack of studies that experimentally examine equivalence and reliability between stationary and mobile means of measuring psychophysiological signals (i.e., HR, HRV, and EDA/GSR). Ollander (2015) used a small sample (*n* = 9) to compare laboratory equipment (Biopac) to a wearable device (Empatica E4 wristband) for detecting psychophysiological stress responses (HR and GSR). The E4 wristband measured mean HR similarly to laboratory equipment, but with a lower sampling frequency (i.e., wearable sampling frequency = 4-64 Hz vs. stationary = 1 kHz). However, mean GSR measurements were not analogous, probably due to differences in assessment sensitivity resulting from measurements occurring from differences in electrode placements (fingers in the case of stationary vs. wrist in the case of wearable).

Another study by Ollander and colleagues (2016), resulted in similar promising results in participants (*n* = 7) who underwent two tasks (stressful task vs. control/stress-free task). Both wearable (E4 wristband) and stationary devices (Biopac) had a good estimation of mean HR and good stress discrimination (stress-related vs. stress-free activity). However, skin conductivity (SC) signals again showed low correlation between the two types of devices. Recently, Ragot and colleagues (2017) compared physiological data from the E4 wristband to those of laboratory equipment (Biopac) during an emotion recognition experiment (using emotional pictures). These authors utilized machine learning models for the analysis of their psychophysiological data and found that the mean responses of physiological data (for cardiac features of HR, AVNN, SDNN, RMSSD, pNN50, LF, HF, RF) yielded from the wearable device, were similar to those of the laboratory equipment. Specifically, correlations between the stationary and wearable devices were high for the average values of cardiac features (i.e., from .50 to .99). However, once again the correlations were low for mean skin conductance level (SCL; *r* = .13). Moreover, emotion recognition (i.e., machine learning algorithms trained both devices to recognize valence and arousal) was found to be similar between the two devices, with an accuracy of 66% for recognizing valence and 70% for recognizing arousal. Heathers (2013) examined HR and HRV as measured via a laboratory PowerLab stationary ECG sampler and a Smartphone Pulse Rate Variability system, and found that there was an accurate approximation of mean HR and HRV features (i.e., RMSSDD, SDNN) between the two types of devices, when participants were at rest, during an attentional task, and during an exercise task.

Overall, existing limited findings present promise for the accurate, reliable and comparable measurement of some psychophysiological indices (e.g., mean HR, HRV indices) via wearable devices compared to traditional stationary ones. However, some signals (e.g., mean values of EDA) present as problematic when ambulatorily assessed, resulting in low reliability. Some measurement reliability problems of wearable devices may be attributed to differences in sampling frequencies or electrode placement locations. More research is needed to examine concordance and discordance between stationary and wearable devices over time and across various conditions (e.g., rest vs. doing a task), and utilizing larger samples of participants. Such research will not only provide necessary information of the use of ambulatory devices at the present time, in daily life and for longer and continuous duration, but will also stimulate further research into new and improved models of such equipment.

Furthermore, with regards to HRV indices specifically, psychophysiological data have been analyzed traditionally with much manual effort, with peaks of the electrocardiography signals (ECG) detected and manually checked for artifacts and noise and then fed into a specific software (e.g., Kubios, AcqKnowledge, Statistica, Artiifact) for extracting HR and HRV features. However, with the potential amount of data able to be harnessed via wearable devices, such traditional analysis will be impossible. Automated analyses using software (e.g., Python programs), have the potential to aid in the accurate analysis of a large volume of collected psychophysiological data (e.g., features of HRV), however, to date they have been largely underutilized in this specific context.

1.1. The Present Study

The purpose of this study was twofold. First, to compare the psychophysiological data -i.e., average HR, HRV features (average RR, SDNN, RMSSD, pNN50) and average SCL- captured with a wearable device (Microsoft Band 2; https://www.microsoft.com/en-us/band; a device different than those used in previous studies) to those of a standard stationary psychophysiological device (i.e., Biopac MP150) in terms of association and congruence between signals across both a resting period (baseline phase) and an experiment of pain induction (i.e., cold pressor task). The Microsoft Band 2 was used for this study's purposes rather than other wearable devices tested in previous studies (e.g., E4) due to its significantly lower cost compared to E4, and its easy extraction and handling of sensor data (fewer limitations imposed by the manufacturers allowing for easier manipulation of sampling frequencies- a limitation identified in previous studies). Also, the Microsoft Band 2 is independent from intermediate cloud-based platforms for extracting the data from the device, which allows for a higher longevity of the device in case of product discontinuation; in fact, we developed our own mobile app for downloading the data from the device with ease.

The second purpose of the study is to compare traditional analysis vs. automated analysis of HRV features. Traditionally, HR and SCL data is analyzed using software such as AcqKnowledge. Regarding HRV however, there is great manual data cleaning of artifacts that is involved which takes time, effort, and reduces the amount of data researchers can collect and analyze (Malik et al., 1996). Thus, we wanted to examine whether we could automate this traditionally manual analysis by developing data processing programs using a simple programming language such as Python. Examining the psychophysiological responses from wearable vs. stationary devices within a pain induction paradigm, can identify whether they can reliably assess emotional responses to stressors (acute pain). Also, it can allow for the comparison of responses captured by the two devices between a resting period and an experimental acute pain induction procedure. To our knowledge, this is the first study with such aims that purports to shed light as to the ability of wearable devices to accurately and reliably assess psychophysiological responses to acute pain, and examine in parallel whether automated analyses (using Python programs) can result in an as accurate analysis of HRV collected data as when traditionally analyzed.

2. Method

2.1. Participants

Participants were 43 students from the University of Cyprus (Table 1). The majority of the participants were female (*n* = 37, 86%), aged 18 to 38 years (*Mage* = 21.37 years, *SD* = 3.72); single (*n* = 42, 97.70%); and possessing a high-school diploma (*n* = 30, 69.80%) or a Bachelors' degree (*n* = 11, 25.60%). All participants were recruited from psychology classes and received course credit for participation. Exclusion criteria consisted of presence of any pain-related disorder, Raynaud's disease, heart disease, high blood pressure, hypertension and diabetes; or taking any medication.

"INSERT TABLE 1 ABOUT HERE"

2.2. Self-reported and Behavioral measures

2.2.1. Demographic Information

Basic demographic information was collected including gender, age, marital status, and ethnicity.

2.2.2. Visual Analogue Scale (VAS)

A 10-point visual analogue scale was used to assess pain intensity after the termination of the pain inducing task (ranging from 0 indicating "no pain" to 10 indicating "worst imaginable pain"). The VAS is frequently used in pain induction experiments as a measure of pain intensity (Kohl, Rief, & Glombiewski, 2012; Moore, Stewart, Barnes-Holmes, Barnes-Holmes, & McGuire, 2015).

2.2.3. Pain Tolerance

Pain tolerance was defined as the total length of time that a participant kept his/her hand immersed in the cold water. A digital stopwatch was used to measure the time in seconds. Such assessment of pain tolerance has been utilized by numerous studies and is considered a reliable method (Keogh et al., 2005; Moore et al., 2015).

2.2.4. Pain Threshold

Pain threshold was established via the total amount of time from immersion until a participant verbally reported pain. A digital stopwatch was used to measure time in seconds (Keogh et al., 2005).

2.3. Physiological Measures

2.3.1. Stationary equipment

Physiological data (i.e., mean HR, HRV and mean SCL) were collected using BIOPAC MP150 for Windows and AcqKnowledge 3.9.0 data acquisition software (Biopac Systems Inc., Santa Barbara, CA). Electrocardiography (ECG) signals were recorded via electrodes placed on each participant's inner forearm. Raw ECG signals were filtered by a BIOPAC ECG100C bioamplifier, which was set to record HR from 40 to 180 beats per minute (BPM). HR was converted to BPM on line. Sampling frequency of ECG signals was set to 1kHz. For the ECG signals, interbeat intervals (IBI; time intervals between heartbeats) were extracted from AcqKnowledge and fed into Artiifact (Kaufmann, Sütterlin, Schulz, & Vögele, 2011) from which four HRV time series measures (mean RR, SDNN, RMSSD, pNN50) were extracted and used in the present study (for description of the features and analysis see HRV analysis section). In regards to SCL signals, they were recorded via two silver/silver chloride (Ag/AgCI) electrodes with Velcro straps placed on each participant's second digit of the index and middle fingers on the non-dominant hand. SCL signals were recorded continuously using a BIOPAC GSR100C transducer amplifier in micro-Siemens (μS) and sampling frequency 250Hz.

2.3.2. Wearable device

The second method used to collect psychophysiological data was a wearable psychophysiological monitor bracelet, namely the "Microsoft band 2" (https://www.microsoft.com/en-us/band). Microsoft band measured similar peripheral psychophysiological indices to those recorded with the stationary device from the wrist of the participant (e.g., mean HR in BPM, mean IBI).

The Microsoft band is a wearable wireless multisensor device designed for real-time data acquisition. It can store over 48 hours of data and is reported by the manufacturers to provide accurate data (https://www.microsoft.com/en-us/band/features). It is equipped with Photoplethysmography (PPG) sensor that measures blood volume pulse (BVP) from which HR, HRV and other cardiovascular features can be derived. This sensor uses optical measurements to record and calculate the volume variance of the blood through time (Allen, 2007). For HR, the PPG sensor extracts an individual's BPM whereas for

HRV, the IBIs are extracted. It should be worth mentioning that the IBIs are reported to be the same with RR intervals extracted from the stationary equipment (Ahmed, Begum, & Islam, 2010). Microsoft band 2 also measured skin resistance through a GSR sensor. Transformations were made to the collected data from skin resistance to SCL with the appropriate formulas (for details see the HR and SCL analyses section), in order to be comparable between the two device-types. The sampling frequency of HR as set by its developers was 1Hz, whereas for SCL, there was a choice between 0.2 Hz and 5Hz; we chose the 0.2 Hz in order to more closely simulate out-of-the-lab conditions of measurement and conserve battery life (higher sampling frequencies result in higher battery consumption). Measurement unit of SCL signals was kohms whereas HR was measured in BPM and HRV in IBI.

For purposes of this study, an Android app was used (Galazis, 2017; see the Supplementary Material Figure 1 for a view of the app; available from the authors upon request). This app allows the researcher to select which of the sensors he/she wants to connect to and receive stream data from. The Microsoft band 2 must be first paired with the smartphone through the Bluetooth connection. The data collected from the Microsoft band 2 was stored on the external SD of the smartphone and saved as CSV comma-separated files and then transferred to a PC for processing.

2.4. Cold Pressor Task (CPT)

A pain task was used to induce physiological sensations that would allow for the comparison of stationary with wearable devices. Pain sensations are experienced by everyone and are associated with increased psychophysiological arousal (Birnie, Noel, Chambers, Von Baeyer, & Fernandez, 2011; Birnie, Petter, Boerner, Noel, & Chambers, 2012). In particular, the Cold Pressor Task (CPT) was selected as the method of inducing pain experimentally, since it is a widely established task which allows for a simple yet effective analogue way to induce and examine physiological reactions in the laboratory allowing for the direct comparison of measures assessed via the two different devices (stationary vs. wearable). Some of its most important advantages over other laboratory-based tasks (Birnie et al., 2011, 2012; Von Baeyer, Piira, Chambers, Trapanotto, & Zeltzer, 2005) include: a) it allows researchers to observe individual differences on pain experience with a minimally threatening nature for the participants (no psychological trauma or tissue damage occurs); b) it is free of potentially confounding variables such as fatigue and anxiety; c) it predicts clinical and real-world outcomes, such as increased risk for developing a chronic health condition; and d) participants have control over their exposure to the stimulus by removing their hand from the water. Overall, it is considered a safe method for inducing acute pain with good reliability and validity (Keogh et al., 2005; Von Baeyer et al., 2005).

The CPT consisted of a plastic cooler, filled with cold water and ice and maintained at a temperature range of 0-2ºC. In order to monitor the water temperature, a KT-300 digital thermometer was used measuring temperatures ranging from −50ºC to +300ºC. A wire screen was also used to separate the ice from the water and keep ice away from the participants' hand. This experiment and the pain induction method received approval by the Cyprus National Bioethics Committee (reference: ΕΕΒΚ/ΕΠ/2019/45) and was in accordance with the ethical guidelines of the International Association for the Study of Pain (Charlton, 1995).

2.5. Procedure

Psychology class instructors provided study information to their students. Interested students contacted the primary author via email. Eligibility was then established, and date and time for the experiment was arranged. When the participants came to the laboratory, they received an explanation of the experiment, provided informed consent, completed the demographics questionnaire, and were then seated in a chair. The skin on the arms and fingers was prepared for fitting the stationary equipment electrodes, where for SCL measurement two electrodes were placed on the non-dominant hand (one on the second digit of the index finger and one in the middle finger). For the HR and HRV measurements, two electrodes were placed on the inner forearm of the non-dominant hand and one electrode was placed on the inner forearm of the dominant hand (stationary measurement). Additionally, participants wore on the non-dominant wrist the wearable device.

After checking the physiological signals of both devices, participants were instructed to listen to some music for five minutes and relax (baseline phase) so as to stabilize the physiological recordings and to familiarize participants with their surroundings. After baseline, participants proceeded to the cold pressor pain task (cooler situated next to their chair) and were instructed to immerse their dominant hand into the cooler. Participants were asked to keep their hand immersed until they could no longer tolerate the pain. Unknown to participants, the maximum duration of immersion was limited to two minutes (this limit was set so as not to create any skin damage; Keogh et al., 2005), at which point participants were asked to remove their hand from the cooler. On completion of the task, all electrodes and the Microsoft band were removed, participants completed the VAS scale and were debriefed.

2.6. Heart Rate Variability (HRV) Scoring and Analysis

Stationary data was analyzed in two ways: traditionally and automatically; whereas, data from the wearable device was analyzed only automatically due to the nature of the data (i.e., stationary equipment collects ECG signals that present with clearly visible peaks vs. Microsoft band 2 collects PPG signals whose wave pattern does not show clear peaks). See Figure 1 for the procedures followed for each device for analyzing HRV.

"INSERT FIGURE 1 ABOUT HERE"

For the *traditional manual analysis* of stationary HRV data, AcqKnowledge 3.9.0 and Artiifact softwares were used. Firstly, data was visually examined for artifacts with the noise being manually corrected in AcqKnowledge. Then, the IBI intervals extracted by AcqKnowledge were fed into Artiifact (http://www.artiifact.de/; Kaufmann, Sütterlin, Schulz & Vögele, 2011) in which irregular RR intervals were deleted and time-domain measures of HRV extracted and saved into Excel. Participants with a large number of deleted artifacts (i.e., 5% of RR intervals), were excluded from analyses (*n* = 11). In particular, the time-domain measures extracted from the stationary device were: 1) Mean RR interval (mean distance of intervals between heartbeats); 2) SDNN (standard deviation of intervals between heartbeats); 3) RMSSD (root mean square of successive differences between adjacent intervals); and 4) pNN50 (proportion of differences greater than 50ms). All time-domain measures were calculated for the 5-minute baseline phase and for the experimental phase (at most two minutes).

The second approach to analyzing both stationary and wearable HRV data, consisted of an *automated analysis using Python*, where a Python program was used to automate the traditional analysis (i.e., bypass manual checks and corrections). Automated analyses of stationary vs. wearable HRV data deviated from each other, on the algorithm used for detecting the R-peaks. Hamilton's R-peak detection algorithm was used for stationary equipment, compared to Microsoft band 2 which uses its own detection algorithm (its name is not reported by its developers). However, although a different R-peak detection algorithm was used in each device, counting peaks cannot be that different across algorithims and a recent study (Reinerman-Jones, Harris, & Watson, 2017) supports that the accuracy of the Microsoft Band 2's R-peak detector algorithm is very close to the Hamilton's algorithm (used in this study for stationary equipment analysis).

For the R-peak detection in the stationary equipment, the Python package BioSPPy (https://biosppy.readthedocs.io/en/stable/) was used to apply Hamilton's ECG R-peak algorithm (Hamilton, 2002). Hamilton's algorithm takes into account the height of the peak, peak position, and maximum derivative (function used for R-peak detection = biosppy.signals.ecg.ecg). At first, the raw signal was filtered to reduce noise and then the filtered signal was given as input to the function "biosppy.signals.ecg.hamilton_segmenter" in order to find the positions of R peaks. The filtering function applies band pass filter to the signal using the cutoff frequencies Hz [3, 45] and the absolute value of the signal was calculated over windows of 80ms. Then, QRS complexes were enhanced and the noise was suppressed. The peaks detected in the filtered signal were classified either as a QRS complex or as noise. The detection threshold was determined using the QRS positive peaks (equivalent to RR) and heights of the noise peaks. Regarding the Microsoft band 2, as discussed above, it provided R-peaks by applying its own internal R-peak detection algorithm. After detecting the peaks in both devices, the given files of the stationary and wearable automated methods were read and processed using the Python program we developed, in order to compute the time-domain measures of HRV (see Figure 1).

2.7. Heart Rate (HR) and Skin Conductance Level (SCL) Scoring and Analyses

Stationary HR and SCL data were analyzed traditionally and automatically, whereas wearable data was analyzed only automatically. Commonly used *traditional analysis* of stationary SCL and HR data, was conducted in AcqKnowledge based on its internal algorithm (name not reported by developers). Then, their mean values (SCL in μS and HR in BPM) were extracted into Excel. For the *automated analysis,* the Acq files of the raw stationary HR and SCL data were read by our Python program, and their mean values were computed. For the wearable device, the HR and SCL raw data was computed internally by the Microsoft band 2 and then their mean values were computed using the same Python program as in the stationary automated analysis. The mean HR (in BPM) and mean SCL (in μS) were calculated for an interval of every 10 seconds for each of the phases, for both the wearable and stationary devices. However, in order to compare skin conductance (G) measured by the stationary device (in μS) with the skin resistance (R) measured by the wearable device (in $kohms$ ($K\Omega$)), the units of the wearable device were converted to conductance (in μS). For this, the formula for electrical resistance and electrical conductance, $G = 1/R$ was used; where G is measured in $Siemens$ (S)and R is measured in Ω. S were then converted to μS, where 1S = $10^6\mu S$.

2.8. Statistical Analyses

Data collected from the self-reported (i.e., demographics, pain intensity) and behavioral measures (i.e., pain tolerance, pain threshold) were coded and analyzed with IBM SPSS v.23. After analyzing and scoring the psychophysiological data from both devices, we conducted Pearson product-moment correlations between the data of Microsoft band 2 and Biopac (for HR, HRV and SCL) as well as between the two ways of analyzing and scoring HRV stationary data (i.e., traditionally vs. automated). The traditional analysis of HR and SCL data uses the AcqKnowledge software (hand scoring of data is not necessary), which utilizes a similar procedure to the Python program. Thus, we expected the correlations between the two types of analyses to be very high and are not reported. Correlations between devices were firstly examined on the mean HR and SCL per 10-second intervals and by phase (relaxation vs. experimental pain induction).

Additionally, because Pearson's correlation may be misleading when comparing two devices (i.e., high correlation might not represent high agreement), the Bland-Atman scatterplots were created to graphically present the agreement between stationary and wearable devices (Bland & Altman, 1986) for each phase and psychophysiological feature (mean HR, mean RR, SDNN, RMSSD, pNN50, mean SCL) in Excel and SPSS. Specifically, in the Bland-Altman scatterplots, the differences between two measurements are plotted (on the vertical axis) against their average values (on the horizontal axis). On the plot, three reference lines are presented (i.e., the 95% upper [+1.96 SD] limit of agreement [LoA], the mean difference between the two measurements [Mean Diff.] and the lower [-1.96 SD] LoA). If there is a perfect agreement between the two methods, the mean difference will be close to 0 (and fall on the solid line). If the values remain between the dashed lines (i.e., between upper and lower LoA), the two methods are considered to be in measurement agreement. The Root mean square errors (RMSE) of mean HR, HRV indices, and mean SCL were also calculated to evaluate the spread of errors between predicted (in this case wearable device) and observed values (stationary traditional and automated analyses methods). Finally, differences on the physiological data (mean HR, HRV indices, mean SCL) between baseline and experimental phases of stationary and wearable devices were examined using paired samples t-tests in SPSS.

3. Results

3.1. Pain Outcomes

Across all participants (*n* = 43), ten (23.30%) tolerated pain for the maximum duration (two minutes), 15 (34.90%) tolerated more than a minute but less than two, and 18 (41.90%) tolerated less than one minute. On average, the total sample tolerated pain for 70.98 seconds (*SD* = 35.36). Pain threshold was 36.02 seconds (*SD*=28.12) and mean pain intensity (VAS scale) 7.05 (*SD* = 1.51).

3.2. Comparison of Psychophysiological Indices Between Stationary and Wearable Devices

3.2.1. Heart Rate (HR)

3.2.1.1. Baseline phase

The average HR based on stationary (traditional) analyses was 80.81 (*SD* = 14.97), on the stationary (automated analysis) was 77.41 (*SD* = 13.30) and on wearable (automated analysis) device 77.01 (*SD* = 12.65; see Figure 2).

"INSERT FIGURE 2 ABOUT HERE"

There were significant correlations (Pearson product-moment correlation coefficients) between the traditional analyses of the stationary and automated analyses of the wearable data for the average HR for all time intervals. In particular, a strong association was observed between stationary (traditional) and wearable (automated) devices for the second intervals from 10 to 300 seconds whereas a moderate association was observed only for the first interval of 0-10 seconds. Similar and even stronger correlations were found for stationary data analyzed automatically in association with wearable (automated analysis) HR data. A strong association was observed between stationary (automated) and wearable (automated) devices for all 10 second intervals. Correlations are presented on Table 2.

"INSERT TABLE 2 ABOUT HERE"

Bland-Altman plots comparing stationary with wearable devices for the baseline phase are shown also in Figure 3.

"INSERT FIGURE 3 ABOUT HERE"

The Bland-Altman plots showed a good agreement between the stationary and wearable devices, with almost all of the participants falling between the 95% LoAs and close to 0. This was further supported with the RMSE error. In particular, the RMSE of HR was low between stationary (when analyzed using both methods) and wearable devices (see Table 3).

"INSERT TABLE 3 ABOUT HERE"

3.2.1.2. Experimental phase

The average HR based on stationary traditional analyses was 88.13 (*SD* = 15.98), on stationary automated analysis was 85.65 (*SD* = 15.15) and on wearable (automated analysis) device was 81.94 (*SD* = 11.73; see Figure 2). Correlations between stationary (traditional analysis) and wearable (automated analysis) devices for mean HR were significant for the majority of the intervals. Strong associations were observed for the intervals of 0-10, 20-30, 80-90, 90-100, 100-110 and 110-120. Moderate relationships were observed for the intervals of 10-20 and 30-40 seconds (see Table 2). Comparable results were also observed when stationary data was analyzed automatically; with strong associations for intervals: 0-10, 10-20, 20-30, 30-40, 70-80, 80-90, 90-100, 100-110 and 110-120 seconds. No significant correlations were observed for the intervals 40-50, 50-60 and 60-70 seconds. Inspection of the Bland-Altman plots (Figure 3) for the experimental phase showed a good agreement between the stationary and wearable devices, with almost all of the participants falling between the 95% LoAs and particularly they were close to 0. Similar results were found based on the RMSE. In particular, low error was observed to the measurement of HR between stationary (when analyzed using both methods) and wearable devices (see Table 3).

3.2.2. Heart Rate Variability (HRV)

3.2.2.1. Baseline phase

The mean values of each time-domain measure for stationary and wearable devices are presented in Figures 4a-d (i.e., mean RR [Figure 4a], SDNN [Figure 4b], RMSSD [Figure 4c] and pNN50 [Figure 4d]).

"INSERT FIGURE 4 ABOUT HERE"

When stationary data was analyzed traditionally, a significantly almost perfect positive correlation was found between stationary (traditional) and wearable (automated) for the mean RR measure (r = .99, p < .001), whereas for the rest of the indices, the correlations were significantly strong: RMSSD (r = .89, p < .001), SDNN (r = .85, p < .001), and pNN50 (r = .83, p < .001). Similar results were found when stationary equipment psychophysiological signals were analyzed automatically: mean RR (r = .99, p < .001), pNN50 (r = .85, p < .001), RMSSD (r = .56, p < .001) and SDNN (r = .55, p < .001). Further inspection of the Bland-Altman plots (Figures 5-8) showed a good agreement for all HRV features between the stationary and wearable devices, with the majority of the participants (n=29; 91%) falling between the 95% LoAs and were close to 0 (for mean RR, SDNN and RMSSD).

"INSERT FIGURES 5-8 ABOUT HERE"

Moreover, inspection of the errors between stationary and wearable devices with RMSE, for mean RR a relatively low error was observed (see Table 3). For SDNN, the RMSE was moderate whereas for RMSSD and pNN50 was high.

When comparing the two ways of analyzing the stationary data (mean values on Figures 4a-d), all correlations of the time-domain measures of HRV were significantly strong: mean RR: r = .99, p < .001; SDNN: r = .63, p < .001; RMSSD: r = .63, p < .001; and pNN50: r = .99, p < .001. These findings were further supported by the Bland-Altman plots which indicated a good agreement between traditional and automated analyses methods for stationary collected HRV data (especially for mean RR, SDNN and

RMSSD), with almost all of the participants falling between the upper and lower 95% LoA, were close to 0 and having low spread (Figures 5-8).

3.2.2.2. Experimental phase

The mean values of each time-domain measure of stationary and wearable devices are shown in Figures 4a-d (i.e., mean RR [Figure 4a], SDNN [Figure 4b], RMSSD [Figure 4c] and pNN50 [Figure 4d]). A significant strong correlation was found between stationary equipment (when analyzed traditionally) and wearable (automated analysis) device for mean RR (r = .74, p < .001) and moderate correlation was observed for pNN50 (r = .40, p < .05). No significant correlation was found between the two devices for SDNN and RMSSD. Comparable results were observed when stationary data was analyzed automatically with significantly strong positive correlations for mean RR (r = .76, p < .001), and pNN50 (r = .53, p < .001) between the two devices. Moderate positive correlation was observed for RMSSD (r = .35, p < .05), whereas no significant correlation was found for SDNN. Further inspection of the Bland-Altman plots (Figures 5-8) showed a good agreement for mean RR between the stationary and wearable devices, with the majority of the participants falling between the 95% LoAs and having low spread. However, for RMSSD, SDNN and pNN50 the agreement between the two devices was weaker. Similarly, when the error between stationary and wearable devices was inspected with RMSE, it was observed to be relatively low for mean RR (stationary [traditional] = 58.09ms; stationary [automated] = 19.03ms). For the rest time-domain measures, the RMSE was high (see Table 3).

Regarding the correlations between traditional and automated ways of analyzing the stationary psychophysiological data, significant strong associations were observed for mean RR (r = .82, p < .001), pNN50 (r = .86, p<.001) and RMSSD (r = .54, p < .01), and a moderate association for SDNN (r = .35, p < .05). The Bland-Altman plots further corroborate these findings and indicate good agreement between the two analyses methods especially for mean RR and SDNN, with almost all of the participants falling between the upper and lower 95% LoA, close to 0 and having low spread (Figures 5-8). However, the agreement between the two ways of analyzing the stationary data for RMSSD and pNN50 was weaker with high spread presented.

3.2.3. Skin Conductance Level (SCL)

3.2.3.1. Baseline phase

The mean SCL (in μS) of stationary (traditional analysis) was 5.20 (*SD* = 2.49), the stationary (automated analysis) 5.20 (*SD* = 2.49) and of the wearable (automated analysis) device was .21 (*SD* = .26; see Figure 9 for each 10 second interval values).

"INSERT FIGURE 9 ABOUT HERE"

No significant correlations were found between stationary (analyzed using both ways) and wearable (automated analysis) for any of the 10sec intervals. Additionally, according to the Bland-Altman plots (Figure 10), the agreement between stationary (analyzed using both ways) and wearable (automated) devices for the baseline phase was observed to be limited. Error was also found to be high (based on RMSE) between stationary and wearable devices (see Table 3).

"INSERT FIGURE 10 ABOUT HERE"

3.2.3.2. Experimental phase

The mean SCL (in μS) of stationary (traditional analysis) was 7.11 (*SD* = 3.33), stationary (automated analysis) 7.11 (*SD* = 3.33) and of the wearable (automated analysis) device was .32 (*SD* = .30). The values of mean SCL data for each 10 second interval of the two devices are shown in Figure 9. The correlations between stationary (analyzed using both ways) and wearable (automated analysis) data, presented with no significant associations. These findings were further supported by the Bland-Altman plots which indicated a limited agreement between stationary (analyzed using both ways) and wearable (automated analysis) devices (Figure 10). High error was also observed (based on RMSE) between stationary and wearable devices (see Table 3).

3.3. Comparison of Physiological Responses between Baseline and Experimental Phases

Paired t-tests comparing physiological responses between baseline and experimental phases for stationary (analyzed using both methods) and wearable (automated analysis) devices, presented with significant increases for mean HR, HRV measurements of SDNN and RMSSD, and mean SCL at the experimental phase (Supplementary Table 2). There was also a significant difference in both devices, between baseline and experimental phase for mean RR, with means decreasing at the experimental phase. For pNN50, significant increases at the experimental phase were observed only in the wearable (automated analysis) device (Supplementary Table 2).

4. Discussion

The primary objective of this study was to establish a methodology and compare congruence of psychophysiological data collected via two types of devices (stationary vs. wearable) in an experimental acute pain induction task. At the baseline phase, significant correlations (and agreement based on Bland-Altman's method) were found between stationary and wearable devices for mean HR and all of the HRV features (i.e., mean RR, pNN50, SDNN, RMSSD). This held mostly for the experimental phase as well, where significant correlations and agreement were observed for mean HR and most of the HRV features (i.e., mean RR, RMSSD). No significant correlations (and worse agreement than the other HRV features) were observed however, for the HRV time-domain measure of SDNN during the experimental phase. Additionally, the RMSE showed high errors between stationary and wearable devices on the measurement of specific time-domain HRV measurements (RMSSD, pNN50) especially during the experimental phase. As a result, findings suggest that when it comes to measuring distances there is equivalency between the two devices, but when it comes to variances and SDs there are differences in the way these are calculated. Additionally, it is advisable to future researchers who would like to use wearable devices for data collection to measure mean HR and mean RR as they showed greater stability and were very reliable especially in absence of motion artifacts (see Table 4 for recommendations for future researchers).

"INSERT TABLE 4 ABOUT HERE"

RMSSD should be preferred to SDNN and pNN50, as it estimates better short-term HRV and is capable of reflecting cardiac vagal tone than the other two time-domain measures. Moreover, according to the Task Force of HRV (Malik et al., 1996) and the recommendations of Laborde and colleagues (2017), SDNN is not a well-defined statistical quantity measure due to its dependence on the duration of the recording period.

In fact, duration of HRV assessment (i.e., maximum of 2 minutes) was one of the main factors contributing to the significantly smaller correlations between stationary and wearable HRV measurements during the experimental pain induction phase. Based on the recommendations of the Task Force of HRV

(Malik et al., 1996) and of Laborde, Mosley and Thayer (2017), HRV should be ideally recorded for the duration of five minutes in order to ensure signal stability. However, in the present study the total duration of the experimental phase was at most two minutes due to the nature of the experiment (i.e., it is impossible and unethical for the participants to have their hand in the cold water for five minutes as skin damage might be caused). Despite the 5-minute recommendation, there are studies that report that recordings of only a few minutes are sufficient for the calculation of certain HRV parameters (Hamilton, Mckechnie, & Macfarlane, 2004; Migliaro, Canetti, Contreras, & Hakas, 2003; Schäfer & Vagedes, 2013) and this is why we decided to proceed with the HRV analysis during the experimental phase. Future studies should however explore differences between wearable and stationary devices in all HRV signals during lengthier emotion and physical sensation induction procedures.

The second factor contributing to small differences in correlations between the assessed signals from the two devices during the experimental phase, may be related to the nature of the experiment. There was some movement involved when participants inserted their arm in the water, and this may have introduced some motion artifacts and noise into the psychophysiological assessments. This was further supported by the RMSE error which was much higher in all physiological measurements during the experimental phase. Indeed, there have been reports that the PPG sensor, used to measure HR and HRV in wearable devices is susceptible to motion noise (Baek & Shin, 2017). As motion increases, correlations decrease due to motion artifacts (Georgiou et al., 2018). This is an important issue that developers of wearable devices need to address, since the purpose of these devices are to be worn while a person goes about their daily life, which involves a multitude of movements. Both wearable devices minimizing motion artifacts and algorithms removing artifacts during intense motions are thus greatly needed (Table 4 for recommendations).

Findings were not very encouraging for SCL measurements. Similar to previous research (Heathers, 2013; Ollander, 2015; Ollander et al., 2016; Ragot et al., 2017), no significant correlations were observed between stationary and wearable devices for the mean SCL for both the baseline and the experimental phases. It is believed that the differences in location from where the data are collected (wrist in wearable vs. fingers in stationary equipment), what has been collected (electrical resistance in wearable vs. electrical conductance in stationary equipment), and the lower sampling frequency of wearables (0.2Hz vs. 250Hz in stationary) that need to conserve battery life compared to stationary equipment that are continuously plugged into a power supply, contribute to these differences. The cost of increasing sampling frequencies is decreased battery life, so this is another problem that wearable developers need to solve so as to improve accuracy. However, the influence of lower sampling frequency is minor based on previous studies (Ollander, 2015; Ollander et al., 2016) and positioning of the measuring sensors is the major contributor to lack of measurement concordance. Measurement of EDA is traditionally assessed in the laboratory via finger placement because the resulting signal resolution is more sensitive and accurate compared to that from other bodily locations (e.g., wrist; Bergstrom, Duda, Hawkins, & McGill, 2014). In particular, based on the introductory guide to EDA of BIOPAC (BIOPAC, n.d.) and findings of previous studies (Ollander, 2015; Ollander et al., 2016), SCL measurements of the wrist and fingers are dissimilar due to the different capacitive properties of the two locations and the higher number of eccrine sudoriferous glands (which regulate EDA) in fingers compared to the wrist. Therefore, this is another challenge for wearable device developers to find a way to more accurately and sensitively measure SCL. Though correlations between stationary and wearable data for SCL were non-significant, showing a possible lack of congruence, SCL as measured via both devices still resulted in fluctuations between the two study phases. This may suggest that though SCL measured via wearables is not exactly equivalent to that of the stationary equipment, it may still be able to detect changes in SCL as a result of being exposed to a stressor (acute pain). Future researchers should also be very careful when assessing skin conductance and prefer wearable devices collecting the same data (skin conductance and not skin resistance) as stationary devices.

This study also examined different ways of analyzing psychophysiological data for HRV, comparing the traditional means of analysis to an automated analysis (the Python program implementing Hamilton’s algorithm in this case). For both phases (baseline and experimental), all correlations between means of analysis for HRV were significant. These were further supported with the Bland-Altman plots which indicated a good agreement between the automated and traditional stationary analyses. Therefore, analyzing stationary HRV data with computer-run algorithms results in similar data as when traditionally analyzed.

4.1. Limitations

This study has several limitations that need to be considered in the interpretation of findings. First, the generalizability of findings to other non-clinical and clinical populations might be limited because this study included only university students. Though this study utilized a larger sample size than previous research (Ollander, 2015; Ollander et al., 2016; Ragot et al., 2017), the sample size may still be limited. Second, this study was conducted in a highly controlled laboratory setting, limiting thus the generalization of the results to real life situations. However, to be able to assess wearable measurement accuracy and validity, this laboratory ground-proof is needed, and the problems identified with measurement (e.g. movement artifacts and noise, electrode placement) need to be resolved.

Third, this study was limited to the calculation only of time-domain measures of HRV. This is due to the limitation that the wearable device has, of providing PPG signals and not raw ECG signals where RR peaks are clearly visible. In addition, the Microsoft band 2 used its own algorithms (names not reported) for analyzing HR, SCL and for detecting R-peaks for HRV. In the present study, the Hamilton’s algorithm was used to analyze the HRV data, which is considered to be very close to the one used by Microsoft band 2 (Reinerman-Jones, Harris, & Watson, 2017). Researchers should be careful when making decisions on the wearable device they choose to use in future studies, to ensure that these devices provide raw signals and allow for the extraction of the data (see Table 4).

Another limitation consists of the sampling frequency of the Microsoft band 2. Specifically, the sampling frequency of HR and HRV was set to 1Hz by the wearable developers and for SCL at 0.2Hz. The configuration used, was one provided by the manufacturers’ options and is the recommended one for daily life wear and the ideal so as to conserve battery life. Sampling frequencies of wearable devices range for SCL from 0.02Hz to 64Hz, whereas for HR and HRV from 1Hz to 100Hz (but typically ranges from 1Hz to 10Hz), placing thus the chosen frequencies within this range (Chen, Hu, & Lin, 2018;

Ollander, 2015). In contrast, stationary equipment do not have such limitations and the Biopac sampling frequency used in research tend to be set to 1kHz for ECG signals and 250Hz for SCL. Though we transformed the sampling frequencies for analysis purposes, the data collected via larger sample frequencies is richer and more sensitive. A solution to this could be the development of analysis techniques that would take into consideration this limitation by using, for example, data explosion methods (Demosthenous, 2019). Wearable developers should consider ways of improving sampling frequencies while extending battery life, so as to make these devices more usable for extended field wear.

4.2. Implications for Researchers and Future Directions

This study is the first to compare psychophysiological data recorded using a wearable device (Microsoft band 2) to standard stationary equipment (Biopac MP150) in an experiment inducing acute physical pain. This is also the first study directly comparing different means of analyzing HRV psychophysiological data (traditional vs. automated means). Therefore, this study consists of an important first step in establishing the methodology for comparing stationary to wearable devices and providing the basis so as to be able to utilize wearables in the accurate and reliable assessment of psychophysiological signals in real-life situations.

Findings showed relative equivalence in HR and HRV between the devices (stationary and wearable). The lower equivalence in some parameters such as SDNN could in theory be remedied by extracting the RR intervals from the ambulatory device, and then feeding them into an algorithm like Artiifact, as is done typically for stationary equipment. However, SCL presented as problematic, as equivalence was not established between the two types of devices, yet differences between the baseline and experimental phases were evident suggesting that the wearable was able to capture differences in SCL but at different levels of measurement. More research is needed in regard to SCL, to be able to establish equivalency and congruence in measurement via wearable devices. Future research should especially focus on finding solutions regarding the measurement of SCL and on establishing sampling frequency equivalency for wearable devices. Further, additional algorithms can be examined or trained in the analysis of psychophysiological data and especially in how to identify and remove noise when assessing HRV. Developers of wearable devices are recommended to provide the raw signals of psychophysiological data, in order to allow for the examination not only of time-domain measures of HRV but also of frequency-domain measures. Once the identified issues are resolved, there can be expansion of the field in the real-time assessment of psychophysiological data within each persons' environment and life situations.

Overall, this study demonstrated the promise of using wearable devices for capturing psychophysiological indices (especially HR and HRV) in an as accurate, reliable, and sensitive way similar to stationary means of measurement. Additionally, algorithmic approaches can be used to analyze the data faster and with less cost and effort on behalf of researchers, allowing for large amounts of data to be recorded and more efficiently analyzed. Establishing wearable data collection as equally reliable, sensitive, and accurate as laboratory stationarily collected data, will open the field in new paths of research and application, and especially lead to the real-time within-a-persons' context assessment of psychophysiological reactions.

5. References

Ahmed, M. U., Begum, S., & Islam, M. S. (2010). Heart rate and inter-beat interval computation to diagnose stress using ECG Sensor Signal. *MRTC Report, 4*.

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, *28*(3), 1–39. https://doi.org/10.1088/0967-3334/28/3/R01

Baek, H. J., & Shin, J. W. (2017). Effect of missing inter-beat interval data on heart rate variability analysis using wrist-worn wearables. *Journal of Medical Systems*, *41*(10). https://doi.org/10.1007/s10916-017-0796-2

Bergstrom, J. R., Duda, S., Hawkins, D., & McGill, M. (2014). Physiological response measurements. In *Eye Tracking in User Experience Design* (pp. 81-108). Morgan Kaufmann. https://www.doi.org/10.1016/B978-0-12-408138-3.00004-2

Birnie, K. A., Noel, M., Chambers, C. T., Von Baeyer, C. L., & Fernandez, C. V. (2011). The cold pressor task: Is it an ethically acceptable pain research method in children? *Journal of Pediatric Psychology*, *36*(10), 1071–1081. https://doi.org/10.1093/jpepsy/jsq092

Birnie, K. A., Petter, M., Boerner, K. E., Noel, M., & Chambers, C. T. (2012). Contemporary use of the cold pressor task in pediatric pain research: A systematic review of methods. *Journal of Pain*, *13*(9), 817–826. https://doi.org/10.1016/j.jpain.2012.06.005

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307-310.

Charlton, E. (1995). Ethical guidelines for pain research in humans. Committee on ethical issues of the International Association for the Study of Pain. *Pain, 63*(3), 277-278.

Chen, H. K., Hu, Y. F., & Lin, S. F. (2018). Methodological considerations in calculating heart rate variability based on wearable device heart rate samples. *Computers in Biology and Medicine*, *102*, 396-401. https://doi.org/10.1016/j.compbiomed.2018.08.023

Demosthenous G. (2019). *Machine learning for the prediction of emotional coping using psychophysiological signals* (Unpublished master thesis). University of Cyprus, Nicosia, Cypris.

Forsyth, L., & Hayes, L. L. (2014). The effects of acceptance of thoughts, mindful awareness of breathing, and spontaneous coping on an experimentally induced pain task. *Psychological Record*, *64*(3), 447–455. https://doi.org/10.1007/s40732-014-0010-6

Galazis, C. (2017). *Non-intrusive physiological wearable devices for identifying individual difference parameters using supervised classification learning algorithms* (Unpublished bachelor thesis). University of Cyprus, Nicosia, Cyprus.

Garbarino, M., Lai, M., Bender, D., Picard, R. W., & Tognetti, S. (2014). Empatica E3-A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. *Proceedings of the 2014 4th International Conference on Wireless Mobile Communication and Healthcare-"Transforming Healthcare Through Innovations in Mobile and Wireless Technologies", MOBIHEALTH 2014*, 39–42. https://doi.org/10.1109/MOBIHEALTH.2014.7015904

Georgiou, K., Larentzakis, A. V., Khamis, N. N., Alsuhaibani, G. I., Alaska, Y. A., & Giallafos, E. J. (2018). Can wearable devices accurately measure heart rate variability? A systematic review. *Folia Medica*, *60*(1), 7–20. https://doi.org/10.2478/folmed-2018-0012

Goncu-berk, G., & Topcuoglu, N. (2017). A healthcare wearable for chronic pain management. Design of a smart glove for rheumatoid arthritis. *The Design Journal*, *20*(1), S1978–S1988. https://doi.org/10.1080/14606925.2017.1352717

Hamilton, P. (2002, September). Open source ECG analysis. In *Computers in Cardiology* (pp. 101-104). IEEE. http://doi.org/10.1109/CIC.2002.1166717

Hamilton, R. M., Mckechnie, P. S., & Macfarlane, P. W. (2004). Can cardiac vagal tone be estimated from the 10-second ECG? *International Journal of Cardiology*, *95*(1), 109–115. https://doi.org/10.1016/j.ijcard.2003.07.005

Hayes, S. C., Bissett, R. T., Zettle, R. D., Cooper, L. E. E. D., & Grundt, A. M. (1999). The impact of acceptance versus control rationales on pain tolerance. *The Psychological Record*, *49*, 33–47. http://dx.doi.org/10.1007/BF03395305

Heathers, J. A. J. (2013). Smartphone-enabled pulse rate variability: An alternative methodology for the collection of heart rate variability in psychophysiological research. *International Journal of Psychophysiology*, *89*(3), 297–304. https://doi.org/10.1016/j.ijpsycho.2013.05.017

Kaufmann, T., Sütterlin, S., Schulz, S. M., & Vögele, C. (2011). ARTiiFACT: A tool for heart rate artifact processing and heart rate variability analysis. *Behavior Research Methods*, *43*(4), 1161–1170. https://doi.org/10.3758/s13428-011-0107-7

Keogh, E., Bond, F. W., Hanmer, R., & Tilston, J. (2005). Comparing acceptance- and control-based coping instructions on the cold-pressor pain experiences of healthy men and women. *European Journal of Pain*, *9*(5), 591–598. https://doi.org/10.1016/j.ejpain.2004.12.005

Kohl, A., Rief, W., & Glombiewski, J. A. (2012). How effective are acceptance strategies ? A meta-analytic review of experimental results. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(4), 988–1001. https://doi.org/10.1016/j.jbtep.2012.03.004

Laborde, S., Mosley, E., & Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research - Recommendations for experiment planning, data analysis, and data reporting. *Frontiers in Psychology*, *8*(FEB), 1–18. https://doi.org/10.3389/fpsyg.2017.00213

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., & Schwartz, P. J. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, *17*(3), 354-381. https://doi.org/10.1093/oxfordjournals.eurheartj.a014868

Masedo, A. I., & Rosa Esteve, M. (2007). Effects of suppression, acceptance and spontaneous coping on pain tolerance, pain intensity and distress. *Behaviour Research and Therapy*, *45*(2), 199–209. https://doi.org/10.1016/j.brat.2006.02.006

Migliaro, E. R., Canetti, R., Contreras, P., & Hakas, M. (2003). Heart rate variability: Short-term studies are as useful as holter to differentiate diabetic patients from healthy subjects. *Annals of Noninvasive Electrocardiology*, *8*(4), 313–320. https://doi.org/10.1046/j.1542-474X.2003.08409.x

Moore, H., Stewart, I., Barnes-Holmes, D., Barnes-Holmes, Y., & McGuire, B. E. (2015). Comparison of acceptance and distraction strategies in coping with experimentally induced pain. *Journal of Pain Research*, *8*, 139–151. https://doi.org/10.2147/JPR.S58559

Ollander, S. (2015). *Wearable sensor data fusion for human stress estimation* (Unpublished doctoral dissertation). Technical University of Linköping University, Sweden.

Ollander, S., Godin, C., Campagne, A., & Charbonnier, S. (2016). A comparison of wearable and stationary sensors for stress detection. *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. http://dx.doi.org/10.1109/SMC.2016.7844917

Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, *9*(21), 1–17. https://doi.org/10.1186/1743-0003-9-21

Ragot, M., Martin, N., Em, S., Pallamin, N., & Diverrez, J. (2017). Emotion recognition using physiological signals: Laboratory vs. wearable sensors. In *International Conference on Applied Human Factors and Ergonomics* (pp. 15–22).

Reis, H. T., & Gosling, S. D. (2010). Social psychological methods outside the laboratory. *Handbook of Social Psychology*, *585,* 82–114. https://doi.org/10.1002/9780470561119.socpsy001003

Reinerman-Jones, L., Harris, J., & Watson, A. (2017). Considerations for using fitness trackers in psychophysiology research. In *International Conference on Human Interface and the Management of Information* (pp. 598-606). Springer, Cham. https://doi.org/10.1007/978-3-319-58521-5_47

Ries, A. J., Touryan, J., Vettel, J., McDowell, K., & Hairston, W. D. (2014). A comparison of electroencephalography signals acquired from conventional and mobile systems. *Journal of Neuroscience and Neuroengineering*, *3*(1), 10–20. https://doi.org/10.1166/jnsne.2014.1092

Schäfer, A., & Vagedes, J. (2013). How accurate is pulse rate variability as an estimate of heart rate variability?: A review on studies comparing photoplethysmographic technology with an electrocardiogram. *International Journal of Cardiology*, *166*(1), 15–29. https://doi.org/10.1016/j.ijcard.2012.03.119

Taj-Eldin, M., Ryan, C., O’flynn, B., & Galvin, P. (2018). A review of wearable solutions for physiological and emotional monitoring for use by people with autism spectrum disorder and their caregivers. *Sensors (Switzerland)*, *18*(12). https://doi.org/10.3390/s18124271

Vissers, G., Heyne, G., Peters, V., & Guerts, J. (2001). The validity of laboratory research in social and behavioral science. *Quality and Quantity*, *35*(2), 129–145. https://doi.org/10.1023/A:1010319117701

Von Baeyer, C. L., Piira, T., Chambers, C. T., Trapanotto, M., & Zeltzer, L. K. (2005). Guidelines for the cold pressor task as an experimental pain stimulus for use with children. *Journal of Pain*, *6*(4), 218–227. https://doi.org/10.1016/j.jpain.2005.01.349

6. Author Notes

6.1. Conflict of Interest

The authors declare no potential conflict of interest.

6.2. Name and Email address for reprints

Maria Karekla, Ph.D., Department of Psychology, University of Cyprus, P.O. Box 20537, Nicosia 1678, Cyprus; TEL: 357 22 892100; mkarekla@ucy.ac.cy

Titles and Captions for Figures

Figure 1. Procedure followed by each method for analyzing the HRV data.

Figure 2. Mean HR (BPM) values of stationary and wearable devices (*N* = 32).

Figure 3. Bland-Altman plots comparing mean HR across two phases: a) baseline and b) experimental, each for two analyses: 1) stationary (automated) vs. wearable (automated analysis); 2) stationary (traditional) vs. wearable (automated analysis).

Figure 4. HRV values of stationary and wearable devices (*N* = 32) for time-domain measures a) Mean RR b) SDNN c) RMSSD and d) pNN50.

Figure 5. Bland-Altman plots comparing mean RR across two phases: a) baseline and b) experimental, each for three analyses: 1) stationary (automated) vs. wearable (automated analysis); 2) stationary (traditional) vs. wearable (automated analysis); 3) stationary (traditional) vs. stationary (automated).

Figure 6. Bland-Altman plots comparing SDNN across two phases: a) baseline and b) experimental, each for three analyses: 1) stationary (automated) vs. wearable (automated analysis); 2) stationary (traditional) vs. wearable (automated analysis); 3) stationary (traditional) vs. stationary (automated).

Figure 7. Bland-Altman plots comparing RMSSD across two phases: a) baseline and b) experimental, each for three analyses: 1) stationary (automated) vs. wearable (automated analysis); 2) stationary (traditional) vs. wearable (automated analysis); 3) stationary (traditional) vs. stationary (automated).

Figure 8. Bland-Altman plots comparing pNN50 across two phases: a) baseline and b) experimental, each for three analyses: 1) stationary (automated) vs. wearable (automated analysis); 2) stationary (traditional) vs. wearable (automated analysis); 3) stationary (traditional) vs. stationary (automated).

Figure 9. Mean SCL values of stationary and wearable devices (*N* = 30).

Figure 10. Bland-Altman plots comparing mean SCL across two phases: a) baseline and b) experimental, for stationary (traditional) vs. wearable (automated analysis).