sensors

MDPI

Article

On the Reliability of Wearable Technology: A Tutorial on Measuring Heart Rate and Heart Rate Variability in the Wild

Veronica Dudarev [1,2,*], Oswald Barral [2], Chuxuan Zhang [1,3], Guy Davis [2] and James T. Enns [1]

[1] Department of Psychology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; chuxuan.zhang@alumni.ubc.ca (C.Z.); jenns@psych.ubc.ca (J.T.E.)
[2] HealthQb Technologies Inc., Vancouver, BC V6K 1B5, Canada; oswald@yourhealthqb.com (O.B.); guy@yourhealthqb.com (G.D.)
[3] Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
* Correspondence: vdudarev@mail.ubc.ca

Abstract: Wearable sensors are quickly making their way into psychophysiological research, as they allow collecting data outside of a laboratory and for an extended period of time. The present tutorial considers fidelity of physiological measurement with wearable sensors, focusing on reliability. We elaborate on why ensuring reliability for wearables is important and offer statistical tools for assessing wearable reliability for between participants and within-participant designs. The framework offered here is illustrated using several brands of commercially available heart rate sensors. Measurement reliability varied across sensors and, more importantly, across the situations tested, and was highest during sleep. Our hope is that by systematically quantifying measurement reliability, researchers will be able to make informed choices about specific wearable devices and measurement procedures that meet their research goals.

Keywords: wearables; measurement reliability; heart rate; heart rate variability

check for updates

Citation: Dudarev, V.; Barral, O.; Zhang, C.; Davis, G.; Enns, J.T. On the Reliability of Wearable Technology: A Tutorial on Measuring Heart Rate and Heart Rate Variability in the Wild. *Sensors* **2023**, *23*, 5863. https://doi.org/10.3390/s23135863

Academic Editor: Ki H. Chon

Received: 14 May 2023
Revised: 18 June 2023
Accepted: 19 June 2023
Published: 24 June 2023

1. Introduction

A bathroom scale is a reliable measure of one's weight, provided one stands still on the scale for several moments. Yet, one is likely to discard the measurement shown by the scale if one is startled by a spider during these moments. Trying to measure cardiac signals with a wearable sensor is similar to trying to measure one's weight while dancing on the scale. The fidelity of the measurement will depend not only on the sensor's accuracy but also on the environmental conditions under which the measurement was taken.

The recent rapid proliferation of wearable sensing technology has been accompanied by many tests of their validity [1–7], usually by examining the correlation of the wearable's signal with that of trusted laboratory equipment [8]. A lot of effort has been put into developing hardware and processing algorithms to ensure high correspondence with benchmark equipment [9–14]; for latest reviews, see [15,16]. What is usually overlooked in these tests is that most laboratory measurement procedures severely limit participants' bodily movements and cognitive activities [17]. In sharp contrast, wearable devices promise to measure the same physiological signals across a wide variety of environments and bodily states. Yet, without further testing, there is no guarantee that wearables will yield accurate measurements in all contexts.

For example, Empatica's electrodermal activity (EDA) measurement showed high agreement with the EDA measurement taken under laboratory conditions. Yet, in a study that measured EDA with Empatica for 20 h per participant in their daily lives, 78% of the measurements were artifacts and no meaningful analysis could be performed with the remaining data [18]. In another study, which aimed to establish the validity of Empatica's HRV measurement against a Holter ECG monitor in 24-h ambulatory monitoring, the reported reliability of Empatica's measurement of heart rate variability (HRV) was lower than that of the Holter device, and the proportion of missing data was higher [19].

The present paper focuses on the question: *when* can measurement from a wearable device be trusted? Our goal is to capitalize on the opportunity that wearable sensors offer to measure physiology under conditions where a benchmark device (e.g., ECG) is not viable. At the same time, we are mindful that measurement reliability is likely to vary across sensors and situations. To help researchers assess this variability, we offer several statistical tools that allow assessing measurement reliability [20] *without referencing a second device*. In Sections 3 and 4, we apply these tools to compare the reliability of sensors against one another and to compare the reliability of a given sensor in different situations.

It is textbook knowledge that measurement fidelity can be decomposed into two components: validity and reliability. Validity denotes measurement *accuracy*—usually determined as a correspondence of measurement to another gold-standard measurement of the same variable. Reliability refers to measurement *precision*—that is, consistency of several measurements taken in the same conditions and/or with the same equipment. It is useful to distinguish between validity and reliability, as can be demonstrated by considering an example of a drunk dart thrower. What does being drunk do: limit accuracy, limit precision, or does it limit both? As shown in Figure 1 (https://conjointly.com/kb/reliability-and-validity/ (accessed on 31 November 2019)), the two are orthogonal. If a drunk lacks accuracy, they behave as in A, being consistently off the mark but with reasonable precision. If they lack precision, they perform most poorly in B, though their average accuracy is still good. If they lack both, it is box C. This example helps understand why low reliability makes accuracy hard to determine. There is too much scatter in the data to estimate the mean with confidence. It also helps understand how high reliability does not guarantee accuracy: the dart thrower may be very precise (reliable) but consistently off the mark.

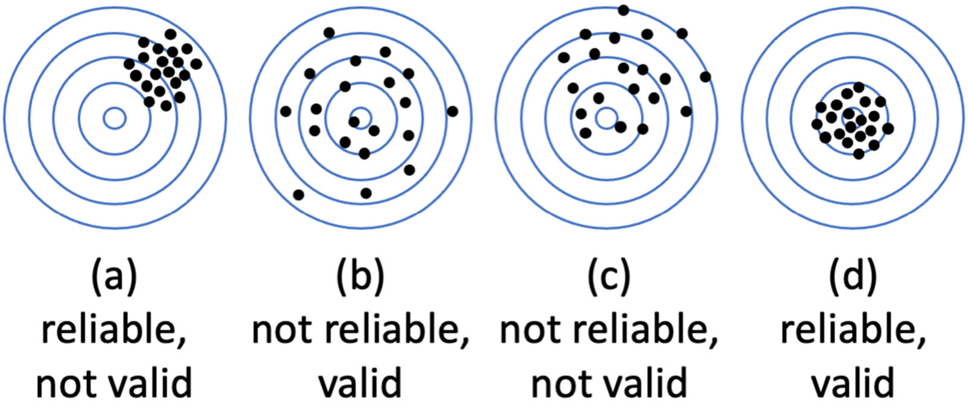

Figure 1. Illustration of the relationships between validity and reliability using the example of drunk dart thrower.

We begin by noting that the theory of measurement reliability was originally developed for assessing fidelity of subjective reports (questionnaires, ratings), not for physiological measures. When participants respond to questionnaires, one cannot fully control the conditions under which they are being filled out (noise, distractions, time pressure) and what factors other than the question may be influencing their responses. Tools to assess response reliability were designed to measure this uncertainty. Psychophysiology borrowed these insights from the theory of reliability to ensure that psychophysiological measurements and associated laboratory protocols resulted in consistent and reliable estimates [17]. Here we perform the same for signals from wearable sensors, while taking into account recent developments in the reliability theory itself [20]. Notice that the reliability analyses we propose are performed on data that are readily available from any wearable device, without additional devices and measurements.

To the best of our knowledge, wearable device reliability has not been considered in detail before. Of the multiple reviews, only one mentioned reliability alongside validity of wearable sensors of cardiac biometrics [21], and referenced only three studies that examined reliability. Kleckner et al. [22] mentioned that for a wearable sensor to be accurate, its measurement has to be reliable, but their proposed framework for choosing a wearable device for research offers no guidelines on assessing its reliability. Here we fill this gap, by offering several readily accessible tools to estimate reliability of measurement with respect to a particular goal.

2. Measuring Reliability

2.1. Definitions

Measurement reliability refers to the consistency of the estimates obtained under hypothetically equivalent conditions; the complement of reliability is measurement uncertainty. Not only does this uncertainty undermine the accurate measurement of a physiological state, but it seriously weakens the ability to use the measured physiological state as a predictor of other outcomes. Spearman [23] noted this issue long ago, showing that low reliability in a predictor variable directly decreases the measured agreement between the predictor and the outcome variables [20,24,25]. Despite this well-known relationship between measurement error and measures of association, many statistical treatments assume that predictor variables (e.g., predictors in a regression) are measured without error, simply because accurate predictor data are the standard assumption in the application of a linear regression [26]. It is also the case that low predictor reliability weakens the results of many statistical tests other than correlations [25]. Although there are ways to assess reliability [20] and to correct some statistics for low reliability [23], many researchers do not use them [20,25]. To summarize, the low reliability of a physiological measurement can prevent one from discovering its true association with other variables.

It is also important to note that not all measurement variability is measurement error. Rather, researchers try to distinguish among sources of potential variance in a measurement, and accordingly, the consequences of these various types of variability on measures of reliability [20,25]. Human physiology is affected by at least two broad groups of factors: constitutional factors and situational variables [27,28]. Depending on the aims of a study, either of these can be considered as noise. For instance, in a study investigating stable differences between individuals, such as differences in personality or physiological traits, situational and state differences between people are a source of noise. Conversely, in studies comparing situational differences, individual differences in personality or physiological traits are a source of noise [29]. Therefore, the type of reliability to be considered depends on the goals of the study.

Constitutional factors refer to enduring or trait-like states of the body. In the case of cardiac measurements, these factors can be intuitively linked to gender, age, body mass index, physical fitness, and chronic medical conditions. This assembly of stable personal traits has predictable influences on blood pressure [30], heart rate [31,32], and heart rate variability [28,33,34]. The stability of a person's physiological parameter measured in different situations is thus referred to as **between-person reliability** (sometimes also as "relative reliability", or parameter level measurement precision—see [8], because it indexes the extent to which a person's parameter is stable relative to other people in different situations. For example, if one's heart rate (HR) or heart rate variability (HRV) is generally high when compared to other people in a laboratory testing, then we would expect their HRV assessed by a wearable device to also show that it was generally higher than other people's data measured by the same device.

The second source of cardiac variability—situational factors—refer to physical activity, stress level, and other more transient physiological states. For example, elevated heart rate along with a reduced HRV is associated with fever [35], inflammation [36], and acute pain [37–40], as well as mental stress [3,41,42] and physical effort [43,44]. It is these situational factors that are spurring much of the current interest in wearable devices. The hope is that tracking users' heart rate biometrics will provide a useful clue for ensuring their health and wellbeing. For instance, studies that compared physically fit people to those who do not exercise as much (between-participant design) tend to show that physical fitness is associated with a higher HRV [44–46]. It is plausible to assume then that increasing one's fitness will result in an increase in HRV, when compared to the same person's HRV before training. However, confirming this conclusion really calls for a within-person study design, i.e., measuring HRV in the same person before and after a change in fitness [47,48]. A within-person comparison of biomarkers therefore calls for the assessment of **within-person reliability.** This is also sometimes referred to as "absolute reliability," and it quantifies the stability of a sensor's readings from the same person in a given situation/state, compared to other states of the same person (for a greater in-depth discussion of between and within-person reliability, see [20].

2.2. Measuring Between-Person Reliability

Between-person reliability refers to the stability of measurement for the same person relative to other people, across time and contexts. In simple words, it is the agreement between measurements taken at different times for the same person (see Figure 2 for illustration). Let us consider a single measurement as

$$x_{ij} = \mu + r_i + v_{ij} \tag{1}$$

where

x_{ij} is a measure taken from individual i in situation j,
μ is the population average,
$\mu + r_i$ is the average for each participant i, and
v_{ij} is measurement error.

Using these terms, between-participant reliability can be expressed as

$$\rho = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_v^2} \tag{2}$$

This is the traditional formulation of population intra-class correlation (ICC–[49]). For a specific sample, in its most general form, it can be computed as:

$$ICC = \frac{MSBS - MSWS}{MSBS + (k-1)MSWS} \tag{3}$$

where

MSBS is the mean sum of squared deviations between the participants,
MSWS is the mean sum of squared deviations within the participants, and
k is the number of measurements (which is required to be equal across participants).

Theoretically, ICC varies from −1 to 1, although in practice, values close to −1 are not usually observed and would suggest the existence of a strong source of systematic variance in the measurement. ICC should be interpreted in the same way as a correlation coefficient: the closer to 1 the higher the agreement, with ICC > 0.75 representing excellent reliability [50].

Several different forms of ICC are available when modeling additional sources of variance. For example, consider a number of patients being examined by several doctors (raters), which is a classic case for ICC application. In this case, it is standard to model the variation among the raters using a two-way ICC, based on the assumption that different people may provide ratings that are systematically different from one another (specific to each rater). For a physiological measure, when there is no a priori reason to assume that individual sensors would be systematically different in their measurement, the most general one-way ICC would be sufficient. The formulas above represent a one-way ICC, capturing between-participant variance (MSBS) against noise variance (MSWS) only. If one wishes to model individual variation along with systematic differences between measurement devices (e.g., different models) or circumstances (e.g., exercise, stress, rest), then a two-way ICC may be applicable.

In addition to deciding whether a one-way or two-way ICC is most appropriate, a decision also needs to be made whether to use single measurement or multiple measurements options. A single measurement ICC estimates representativeness of a single measurement for the person's parameter value. A multiple measurements ICC estimates representativeness of the average across all measurements. For additional details and guidelines, see [20,51].

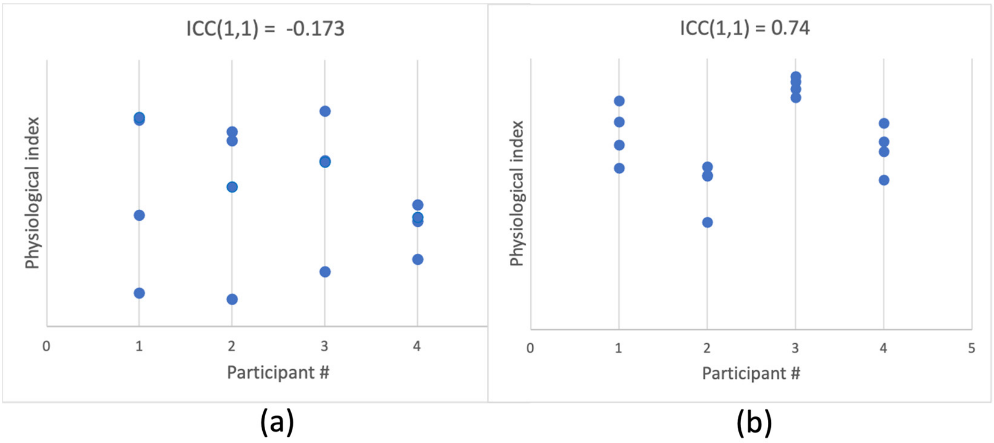

Figure 2. Simulated data of 4 measurements per each of 4 participants. **Panel** (**a**) shows an example of data where participants are hard to distinguish, i.e., between-participant reliability is low, as indicated by a low one-way single-measure ICC. **Panel** (**b**) shows a case in which each participant's physiological index is highly individual, resulting in high ICC and high between-participant reliability.

2.3. Measuring Within-Person Reliability

Within-person reliability refers to the stability of a measurement taken in the same situation, for a given person. In the extreme, it would refer to two identical devices producing equal measurements when used simultaneously on the same person. For a single device, considering the measurement as

$$x_{ij} = \mu + r_i + s_j + v_{ijk} \tag{4}$$

where

x_{ij} is a measure taken from individual i in situation j,
μ is the population average,
$\mu + r_i$ is the average for individual i,
$\mu + s_j$ is the average for a situation j, and
v_{ijk} represents the difference between multiple measurements taken for individual i during situation j.

Here, within-participant reliability would be the opposite of the magnitude of v_{jk}. The conceptual distinction between signal and noise is crucial when considering within-person reliability and will depend on a particular design and aims of the measurement situation (for further guidance, see [20]).

Let us consider a simple example. Imagine we are measuring the heart rate of a particular person, taking several measurements during rest, a few more measurements during mentally stressful activity, and a few more during physical exercise. We would expect high agreement within rest measurements, and distinct but clustered measurements during each type of activity. A mixed model regression that predicts heart rate measured at time 1 from heart rate measured at time 2, with situation (rest, mental activity, physical activity, recovery) as the random factor takes the following form:

$$HR_y \sim a + beta * HR_x + 1|situation \tag{5}$$

As in a general linear regression approach, *beta* is the estimate of the association between the predictor and the predicted variable. In this case, *beta* quantifies the amount of agreement between measurements taken during each situation. If we test more than one person, we should add a random factor of participant to remove the variance associated with individual differences and focus on within-participant consistency:

$$HR_y \sim a + beta * HR_x + 1|situation + 1|participant \tag{6}$$

Figure 3a shows example data simulating four participants in four types of situations where both between- and within-participant reliability (consistency) are high.

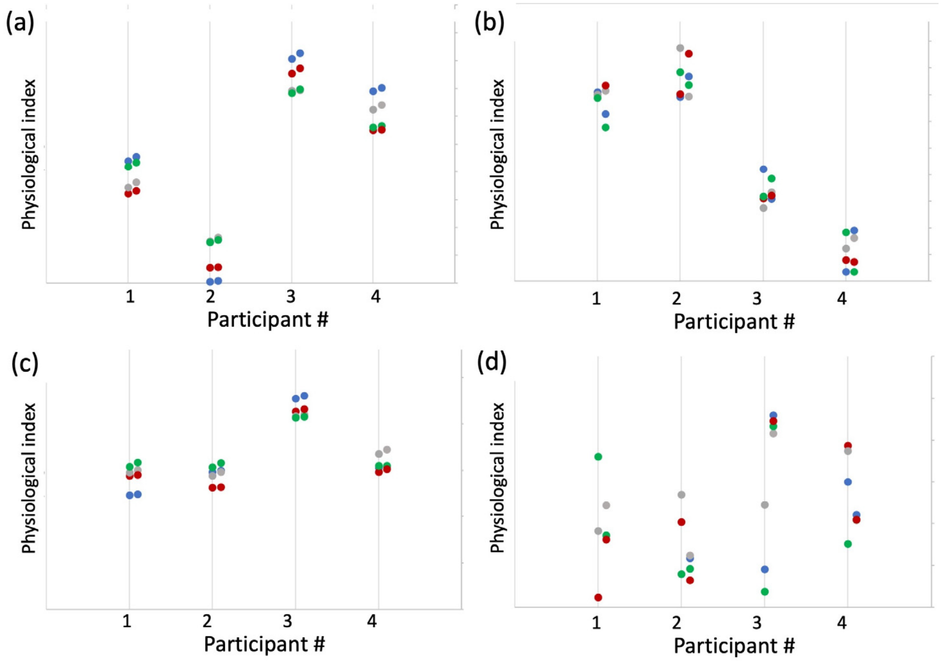

Figure 3. Simulated data of 8 measurements for each of 4 participants, in 4 situations (color coded). **Panel** (**a**) shows an example of high between- and within-participant reliability, where the observations are consistent per participant and per situation. **Panel** (**b**) shows an example of high between-participant reliability yet low within-participant reliability. **Panel** (**c**) shows high within- and low between-participant reliability, where datapoints are consistent within a situation, yet do not reliably distinguish between different individuals. **Panel** (**d**) shows low within- and between-participant reliability.

As can be seen, this type of analysis requires two measurements for the physiological index of interest in each situation. Traditionally, with subjective responses to questionnaires, the questions were divided into two subsamples by the order of their appearance, taking either first and second half of the questionnaire as the two subsamples, or odd vs. even questions (the so-called split-half approach, [23,52]). Unlike questionnaires, physiological measurements, especially those obtained with a wearable device in ecological settings (i.e., daily wear), are performed over a longer time span than is required to fill out a questionnaire, often with the aim of quantifying a change in the person's state from one measurement to the next. Given the volatility of physiological measurement in time, the closer the two samples occur in time, the more similar we would expect them to be. In other words, it is reasonable to expect that time is a systematic factor that must be taken into account. Assigning each datapoint a number by order of its acquisition, then aggregating (averaging) all odd and even datapoints allows one to measure consistency between instances that are taken **as close in time as possible**. We will refer to this as time-sensitive sampling. An alternative approach that is not time-sensitive might involve dividing all measurements into two subsamples randomly. With just one instance of such division, there is a non-zero chance that by coincidence the two subsamples will be uncharacteristically similar or dissimilar. However, if a random split is performed multiple times, we can estimate within-participant reliability from the resulting distribution of *betas*. In the next section, we will compare these two methods of dividing the datapoints to provide an assessment of within-person reliability.

2.4. Empirical Examples

We first apply the approach proposed here to a case of a commercially available PPG-based sensor of cardiac biometrics (heart rate and heart rate variability). The main aim is to demonstrate the degree to which the measurement reliability of a wearable sensor varies with the conditions of everyday life. We focus on the most discussed factors that affect measurement fidelity of wearable sensors: (1) the make of the sensor (hardware + software), and (2) physical activity of the user [1,53]. We achieve this by analyzing a publicly available dataset that contains heart rate recorded by six commercially available PPG-based sensors [54]. We then move to apply the framework proposed here to data collected outside of the laboratory, where comparison to benchmark ECG is not viable. Naturalistic data were collected from 10 healthy participants who were wearing another commercially available PPG sensor, Biostrap, for a week. We compare the reliability of data acquired during sleep and during active wakefulness. In addition, to demonstrate the relationship between the reliability of a measurement and the magnitude of its correlation with another variable [20,23], we test the correlation between the two biometrics explored here with a measure of the participants' mood.

3. Study 1: Reliability of Six Wearable Sensors of Cardiac Biometrics

Bent et al. [2] conducted a study comparing six different wearable devices against ECG to determine measurement fidelity of these devices under different conditions. The data (beats per minute from each device) are publicly available [54]. Here we analyze this dataset by applying the estimation procedures already described to assess the measurement reliability of the devices *without referencing ECG*. We compute between- and within-participant reliability for (a) the six wearable devices, and (b) different activities.

3.1. Method

A total of 53 participants were tested with 6 wearable sensors while engaging in 4 types of activities: rest, paced breathing, physical activity (walking), and typing. In between these activities, the sensor was on the participant's wrist and still recording, and so we include the rest period as the 5^{h} type of activity: task transition. Participants were wearing one or two devices at a time, repeating the activities several times. Because of missing data, some analyses included differing numbers of participants, ranging from 45–53.

3.2. Results

3.2.1. Data Processing

Heart rate was measured as beats per minute (BPM). Of these, we removed all the 0 values, and then values more than 2 standard deviations above or below each participant's average. For between-participant reliability, BPM datapoints were averaged per device and activity type for each participant.

3.2.2. Between-Participant Reliability

A one-way random single-measure ICC(1,1) was computed for the 44–53 participants' mean BPM, separately for each device, with the 5 activity types as the different measurement instances. Table 1 and Figure 4 show the results. The 6 devices appear to have unequal between-participant reliability, Biovotion showing the highest, and generally good reliability of 0.65 (but also the largest number of participants without data), Empatica and Miband showing only fair reliability of 0.38 and 0.37, respectively, and the other three devices showing good reliability [50].

Table 1. Between-participant reliability of HR for the 6 wearable sensors, listed in alphabetical order. Number of participants (N) varies per sensor because of missing data.

Device	ICC(1,1)	95% CI	N
Apple Watch	0.54	[0.44 0.65]	53
Biovotion	0.65	[0.55 0.75]	45
Empatica	0.38	[0.27 0.496]	53
Garmin	0.50	[0.397 0.61]	52
Fitbit	0.48	[0.37 0.59]	53
Miband	0.37	[0.27 0.496]	50

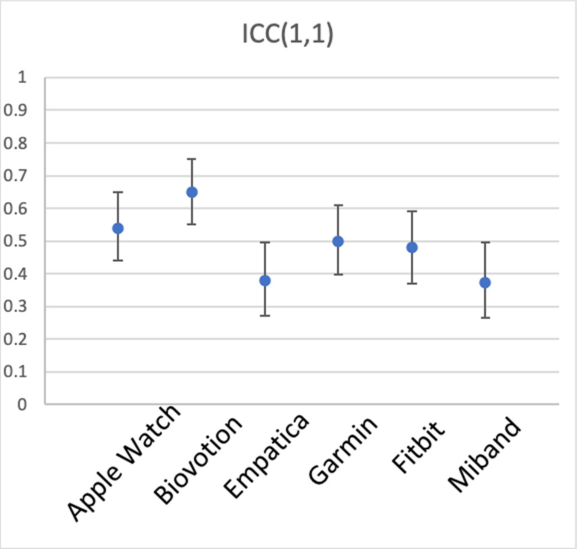

Figure 4. Between-participant reliability of HR for the 6 brands of wearable devices. ICC(1,1) is shown, error bars represent 95% CI.

We then explored whether different conditions of measurement—in this case, different activities—produce data that are more or less representative of individual participants' heart rate (between-participant reliability). To this end, we computed ICC(1,1) for the 5 types of activity, with devices serving as measurement instances. Table 2 and Figure 5 show the results. Breathing, transitioning between activities, and rest elicited the most reliable measurements across devices (ICC(1,1) of 0.66, 0.597, and 0.54, respectively). Physical activity (walking) and typing produced ICC(1,1) of just 0.37, suggesting that measurement was quite noisy during these activities.

Table 2. Between-participant reliability of HR for the 5 activities, listed in alphabetical order.

Activity	ICC(1,1)	95% CI	N
Breathing	0.66	[0.57 0.74]	53
Physical activity	0.37	[0.27 0.48]	53
Rest	0.54	[0.44 0.64]	53
Task Transition	0.60	[0.51 0.69]	53
Typing	0.37	[0.27 0.48]	53

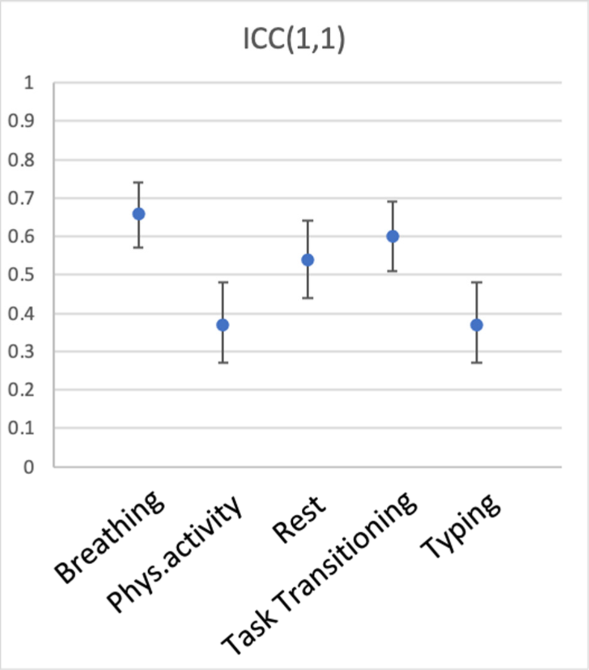

Figure 5. Between-participant reliability of HR across wearable devices for the 5 types of activity. ICC(1,1) is shown, error bars represent 95% CI.

3.2.3. Interim Discussion

Between-participant reliability was examined in the dataset containing heart rate of 53 participants measured with six devices during five types of activities. Reliability varied between the six devices, with Biovotion and the Apple Watch showing highest reliability, closely followed by Garmin and Fitbit, with Empatica and Miband showing lower reliability. Interestingly, a comparison to ECG measurements reported in Bent et al. [2] revealed that the deviation from ECG was lowest for the Apple Watch, followed by Garmin and Fitbit, followed by Empatica and Miband, followed by Biovotion. That is, measurement fidelity as assessed by comparison to ECG (validity) and as assessed by between-participant reliability of the measurement itself (reliability) match closely, with Biovotion being the only exception. It is textbook knowledge that reliability is a necessary, but not sufficient condition for validity of measurement. Thus, this is exactly what the data show for the Apple Watch and Miband. As the reliability of wearable devices decreases across brands, their validity decreases as well. Note too that Biovotion is a clear example of a device that is reliable (not much internal noise), yet not valid (does not correspond to a benchmark device). Thus, high reliability does not guarantee high validity (see Figure 1). However, it is also true that low reliability makes it difficult to determine validity at all. However, once validity is established, is it possible that measurements are unreliable? We addressed this question by investigating different participant activities.

Participants' activities generally affected measurement reliability as expected, with calmer states (breathing, transitioning, rest) producing higher reliability than more intense activities (walking, typing). This is consistent with multiple previous studies, including [2], who reported higher reliability for measurements taken during rest, and reduced reliability with increased levels of activity.

3.2.4. Within-Participant Reliability

Within-participant reliability was assessed using a split-half approach and a mixed model regression. We could not use a time-sensitive approach because time stamps were not available in the dataset; we therefore used a random split-half approach. To explore within-participant reliability of the six devices, we split each participant's heart rate datapoints during each activity into random halves 1000 times, computing a mixed-model regression with the participant as the random factor each time. Reliability was estimated as the average *beta* across the 1000 iterations. To explore the reliability of measurement during different activities, we split each participant's heart rate as measured with each device into random halves 1000 times, submitting that to the mixed-model regression every time.

Figure 6 and Table 3 show within-participant reliability for the six devices. As can be seen from the table, all devices had excellent reliability in measuring heart rate across different situations within a participant. It can also be seen that the reliability of Fitbit was noticeably lower, although still very high.

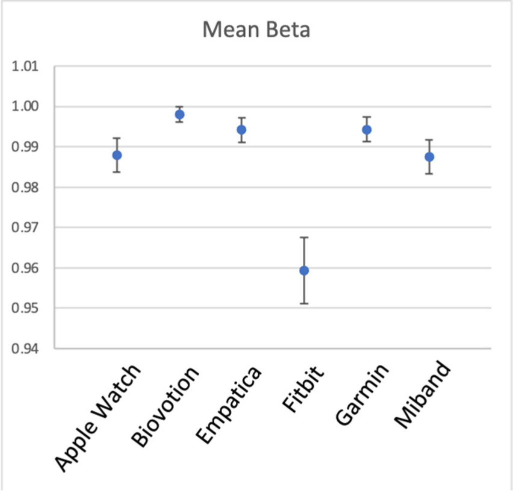

Figure 6. Within-participant reliability of HR for the 6 brands of wearable devices. Error bars represent 1 SD.

Table 3. Within-participant reliability of HR for the 6 wearable sensors, listed in alphabetical order.

Device	Mean Beta	SD of Beta Across Iterations
Apple Watch	0.988	0.008
Biovotion	0.998	0.004
Empatica	0.994	0.006
Fitbit	0.959	0.016
Garmin	0.994	0.006
Miband	0.988	0.009

Figure 7 and Table 4 show within-participant reliability for the five activities. It was also very high across the activities, yet transitioning was noticeably less consistent than the other activities.

3.3. Discussion

We explored within- and between-participant reliability of heart rate measured with six wrist-worn devices during five activities. This demonstration showed that measurement reliability can be estimated without referencing any benchmark device, from the data of a single sensor. We observed noticeable differences in between-participant reliability for the six brands of wearable sensors, and for the different levels of activity participants engaged in. With regard to the two components of measurement fidelity—reliability and validity—the data complied with textbook expectations, showing that high reliability is a necessary, yet not sufficient condition for validity. It showed that validity cannot be inferred from reliability, and that validation of a device is a necessary first step to ensure measurement fidelity under ideal (laboratory) conditions. Yet, as the analysis of different activity levels showed, even once an acceptable level of validity is established under resting conditions, a wearable device can produce a measurement of suboptimal reliability under more active everyday conditions.

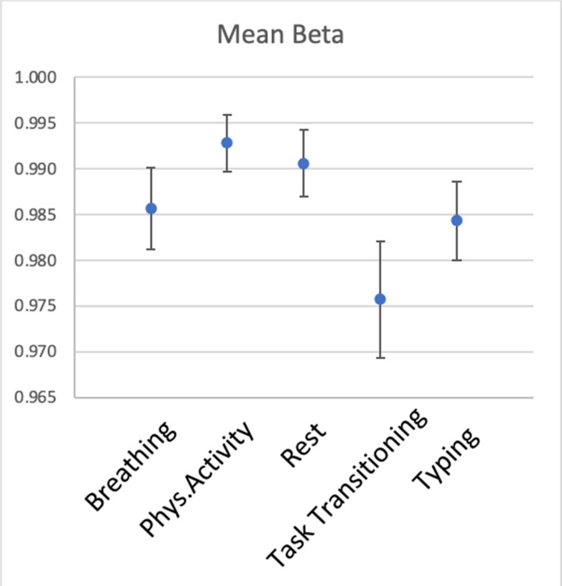

Figure 7. Within-participant reliability of HR for the 5 types of activity. Error bars represent 1 SD.

Table 4. Within-participant reliability of HR of the 5 types of activity.

Activity	Mean Beta	SD of Beta Across Iterations
Breathing	0.986	0.009
Phys. Activity	0.993	0.006
Rest	0.991	0.007
Transitioning	0.976	0.013
Typing	0.984	0.009

Within-participant reliability was very high across devices and activity levels. Heart rate is a great example of a measurement that is highly consistent within participants (high within-participant reliability), but not always acceptable for distinguishing between participants (moderate between-participant reliability). This most probably reflects the nature of heart rate, which has stable and quite narrow limits for a given person, especially during wakeful time. In our next example (Study 2), we examined a less constrained measure—heart rate variability—in order to see how within- and between-person reliability is manifested in this measure.

4. Study 2: Reliability of Biostrap during Sleep and during Wakeful Time

In this study, we tested a commercially available Biostrap wristband sensor for both between-person and within-person reliability of HR and HRV. In our treatment of within-person reliability, we focused on comparing two diurnal states of the user: wakefulness and sleep. Sleep corresponds to time passing with little to no change in the external environment and fewer physiological changes than during wakeful periods (e.g., relatively little physical effort, no eating or talking, relatively little stress and mental effort). We hypothesized that sleep periods would produce less variable and therefore more reliable biometric recordings.

Ten participants wore a Biostrap device continuously for one week. They were instructed to wear the device on their wrist at all times, except when charging the device (about 1 h daily) and when taking a bath or shower.

4.1. Method

4.1.1. Participants

Ten participants (1 male) were recruited through Reservax (https://www.reservax.com (accessed on 31 November 2019)), an online recruitment platform for behavioral studies. The inclusion criteria were: participants at least 18 years of age, without known heart problems or disease, in generally good health, and fluent in written and spoken English. All participants provided informed consent prior to participation. Participants were paid a maximum of $100 CAD for participation, based on their compliance with study procedures. All 10 participants received full payment.

4.1.2. Apparatus

The Biostrap wristband is a commercially available PPG sensor of heart rate (https://biostrap.com (accessed on 31 November 2019)). Biostrap (formerly Wavelet) has been validated against clinical-grade wearable devices and ECG [7,9,55]. The device uses long wavelength light (red) to detect pulse. Automatic sampling is performed once every 5 min (in enhanced mode), each recording lasting for 45 s at 43 Hz frequency. The raw data are stored on the sensor's internal memory, then transmitted to a smartphone app via Bluetooth connection, and then to the Biostrap server where the data are processed.

The output provided by Biostrap includes heart rate in beats per minute (BPM), heart rate variability (HRV) indexed as the root mean square difference between successive heartbeats (rMSSD), oxygen saturation, and respiration rate. This information is provided for each sampled measurement, which can be as frequent as once in every 5 min. The sensor also includes an accelerometer, which provides information on the number of steps completed by the wearer. Based on a combination of these metrics, sleep onset and offset are detected.

The commercial Biostrap smartphone app ordinarily shows the user their heart rate and heart rate variability, number of steps, and a sleep score on the app's home screen (these metrics are shown by default). It also indicates the battery status and the last time the data were synchronized with the app. In this study the app was blinded to participants, so that it was unable to display any biometrics; only the battery status was visible to them. Ecological momentary assessments were delivered using the Ipromptu smartphone app (http://www.ipromptu.net).

4.1.3. Procedure

Invited participants arrived at the lab in the Department of Psychology at UBC for an introductory session, where they were introduced to the Biostrap device, provided personal demographic information, and completed questionnaires on emotional, self-control, and personality traits (which are not reported here).

Each participant received a fully-charged Biostrap wristband to wear for the duration of the study along with a charging plate. The Biostrap app was installed on participants' smartphones and they were instructed on the use of the sensor and how to ensure data were synchronized regularly. Instructions to participants emphasized that they were to wear the device at all times, including times of exercise and sleep, except for when charging the device or taking a shower or bath. Participants wore the Biostrap continuously for 8–11 days.

In addition, participants were asked to track their emotional state using an ecological momentary assessment (EMA) approach. Ipromptu app (http://www.ipromptu.net (accessed on 13 September 2022).) was used to deliver short surveys 6 times a day, at random times between 8 am and 8 pm. If not responded, a prompt repeated twice, with 15-min intervals, and was available for response for several hours. Participants were instructed to respond to at least 1 and as many prompts as they could. On each prompt, an 8-question survey asked participants to rate, on a scale from 1 to 10, how happy/energetic/nervous/afraid/irritable/angry they were and how much pain and discomfort they were feeling, in random order.

Participants returned to the lab at least 8 days after their introductory meeting to conclude the study. They returned the Biostrap devices, were debriefed about the purpose of the study, and paid for their participation.

The study procedures were approved by the institutional Research Ethics Board (approval number H19-01197). Data, materials, and analysis code for this study are available at https://zenodo.org/badge/latestdoi/520639317 (accessed on 13 September 2022).

4.2. Results

4.2.1. Data Processing

Heart rate measurements consisted of raw PPG waveforms, which were processed by Biostrap's algorithms in their servers [9]. The data presented here were based on the aggregated metrics provided by the Biostrap for each successful sample, which included beats per minute (BPM) and heart rate variability (HRV), calculated as the root mean square difference between successive heartbeats (rMSSD). Hereafter, we will refer to HRV for simplicity, instead of rMSSD. These heart rate measures were further screened for artifacts and anomalies in two steps. First, all 0 values were removed (affecting 0% of BPM and an average of 19.26% of HRV samples across all participants). Second, values exceeding each participant's mean by more than 2 standard deviations over the whole observation period were removed (affecting 2.84% of BPM, and 2.19% of HRV across all participants).

Participants' state (asleep vs. awake) was established using the heart rate indices in the following way. We found periods of at least 2 h in duration when BPM samples were successfully recorded at least every 15 min. We then chose the longest such period on each day and assumed that it corresponded to sleep. Although time of day was not a criterion for determining sleep periods, all the sleep periods established in this way happened to occur between 9 pm and 11 am. These criteria allowed us to detect at least 5 periods of sleep for 8 of 10 participants. We recognize that these criteria do not guarantee that participants were awake at all other times, and as such, that this potentially biases awake observations to appear to be more similar to sleep periods. However, to anticipate the results, the density and reliability of HR and HRV assessment during sleep periods defined in this way were greater by orders of magnitude than they were during the defined wakeful times.

Heart rate data were successfully recorded for only 2 sleep periods for one participant and only 1 sleep period for another, and so their data were not included in the analyses. Days with only one HRV sample during wakeful times (5 periods across participants) were also excluded from the analyses. All participants had more than one HRV sample during sleep. These exclusions left us with 8 participants tracked continuously for 5 to 11 days, and a total of 6840 samples for BPM and 5530 samples for HRV.

4.2.2. Descriptive Statistics

Figure 8 shows the frequency of successful heart rate samples for BPM and HRV made during wakeful and sleeping periods. The pattern of these two variables was generally consistent across participants. The mean number of BPM samples acquired for waking periods was 19.42 (SD = 13.52), and the mean number of sleep samples was 74.6 (SD = 24.43). The mean number of HRV wakeful samples was 13.65 (SD = 9.63) and the mean of sleep samples was 69.70 (SD = 22.63).

Figure 9 shows the mean BPM and HRV for each participant, separately for wakefulness and sleep. The figure shows that there are pronounced individual differences in both biometrics, with some participants having consistently higher HRV or BPM than others. The variability of the wakeful measurements is also visibly larger than the variability of sleep measurements. These observations were confirmed by the following analyses.

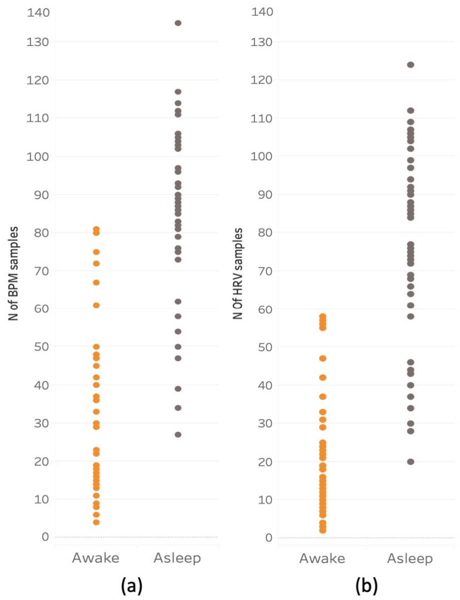

Figure 8. Average number of successful measurements of BPM and HRV per participant per day during wakefulness and sleep. **Panel** (**a**) shows average number of successful measurements of BPM, **panel** (**b**) shows the same data for HRV per day and per participant during wakefulness (orange) and sleep (grey).

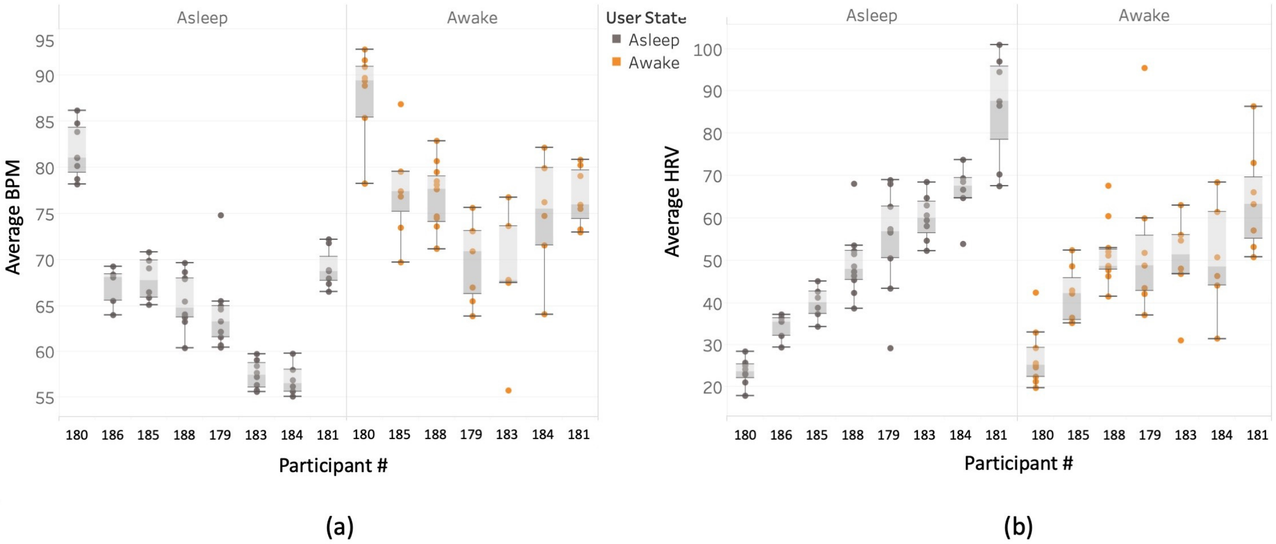

Figure 9. Average HR and HRV per participant during sleep and wakefulness. **Panel** (**a**): The mean BPM for each participant, separated for sleep and wakefulness. **Panel** (**b**): The mean HRV for each participant, separated for sleep and wakefulness. Participants are rank ordered in each panel based on their HRV during sleep. Participant 186 had no biometric recordings for wakeful time.

4.2.3. Between-Participant Reliability

A one-way random single-measure ICC(1,1) was computed for the 8 participants' mean BPM and HRV, separately for wakeful and sleep period samples. ICC values were consistently higher for sleep periods than for wakeful periods. This was true for both BPM values (sleep: *ICC(1,1)* = 0.89.6, 95% CI [0.79 0.97], $p < 0.001$; wakefulness: *ICC(1,1)* = 0.55, 95% CI [0.34 0.83], $p < 0.001$) and for HRV values (sleep: *ICC(1,1)* = 0.84, 95% CI [0.70 0.95], $p < 0.001$; wakeful: *ICC(1,1)* = 0.39, 95% CI [0.19 0.73], $p < 0.001$). These high ICC values for sleep, along with only moderate ICC values for wakefulness imply that individual differences in heart rate and heart rate variability can be measured more reliably with a commercial PPG sensor during sleep than wakefulness.

These data suggest that BPM and HRV measured through a commercial wearable device are relatively stable between people, meaning that a person whose BPM or HRV is higher than other people's on one day/night is likely to have BPM or HRV higher than other people on any other day/night.

4.2.4. Within-Participant Reliability

Within-participant reliability was assessed using a split-half approach and a mixed model regression. We compare two methods of splitting the data: time-sensitive (split into odd and even samples, by order of measurement) and random (dividing the datapoints into two samples randomly, so that a sample from early in the day is equally likely to be paired with a sample from later or earlier in the day). For both methods, a mixed model regression is then computed predicting one estimate of the biometric (e.g., average of the odd datapoints) from the other estimate (e.g., average of the even datapoints), with the participant as the random factor (see Equation (6)), and Satterthwaite's correction for the degrees of freedom.

The time-sensitive method resulted in estimates of the within-participant reliability of BPM and HRV illustrated in Figure 10. Panel a shows that reliability of BPM was very high for the sleep and wakeful periods alike. The effect of predictor BPM was highly significant in both models, *beta* = 0.99, $t(9.17) = 91.8$, $p < 0.001$ and *beta* = 0.82, $t(18.8) = 7.91$, $p < 0.001$, respectively. Panel b shows that the fit between predictor and criterion for HRV was also generally high during sleep, *beta* = 0.96, $t(57) = 51.18$, $p < 0.001$, but not during wakefulness, *beta* = 0.097, $t(47.26) = 0.92$, $p = 0.36$.

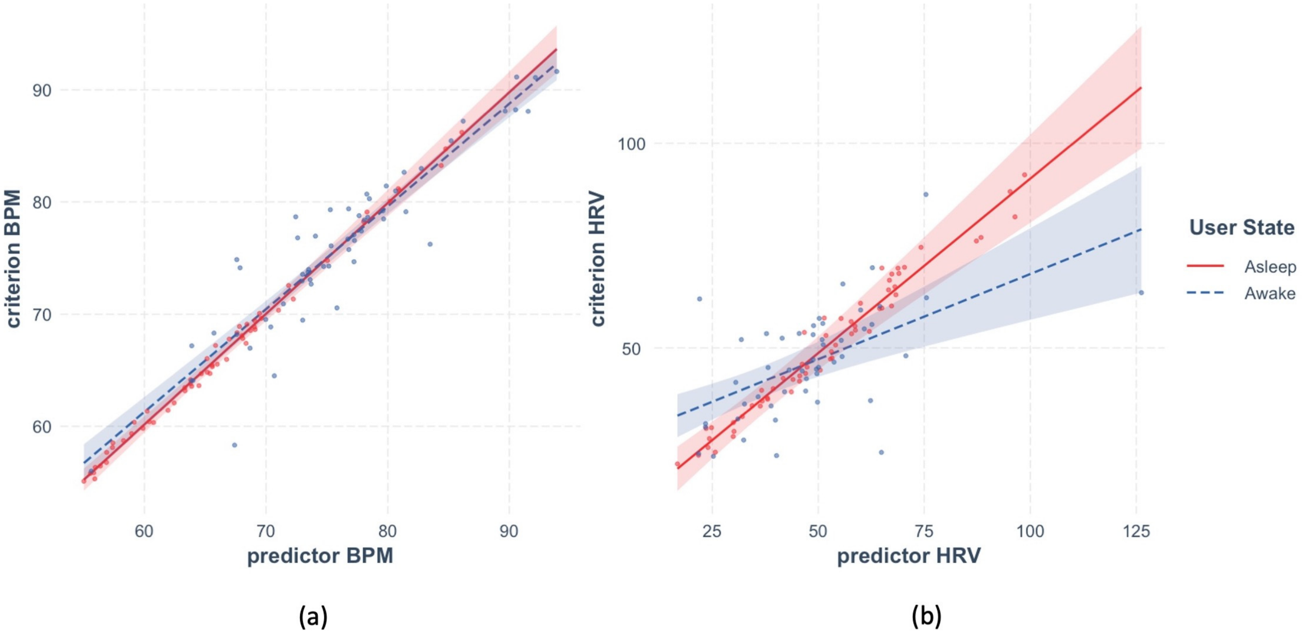

Figure 10. Within-participant reliability of HR and HRV during sleep and wakefulness estimated with the time-sensitive method. **Panel** (**a**) shows within-participant reliability of BPM, **panel** (**b**) of HRV. Data recorded during sleep is shown in red, during wakefulness in blue. Shaded area represents 95% CI, dots represent partial residuals.

Random splitting into the subsamples, as mentioned above, can result in extraordinarily low or high estimates of reliability. Therefore, we performed the split 1000 times, computing a mixed-model regression each time. Figure 11 shows distributions of the resulting *beta* values for HR and HRV during sleep and wakefulness. We estimated reliability as the average *beta* across the 1000 iterations.

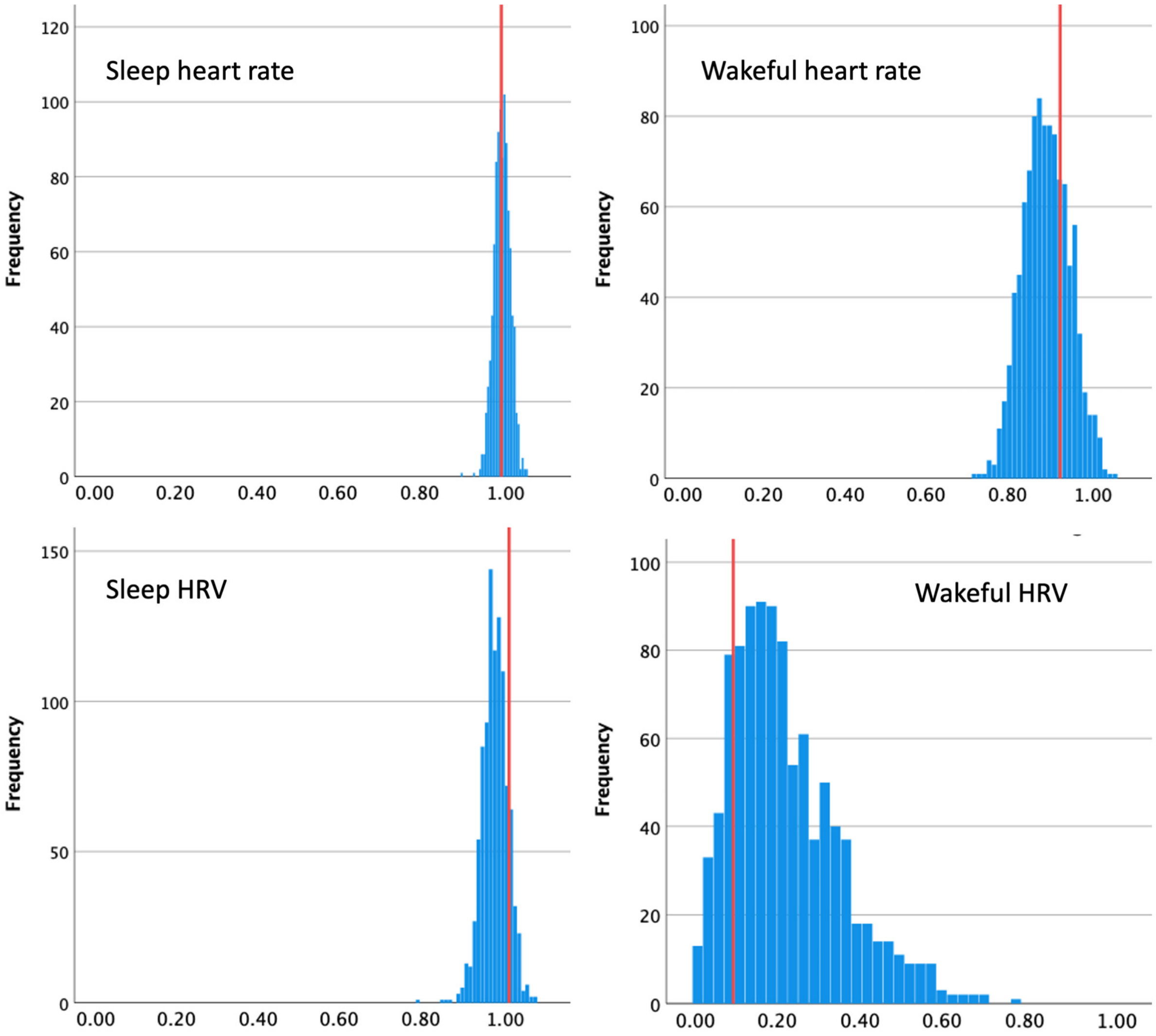

Figure 11. Within-participant reliability of HR and HRV during sleep and wakefulness estimated with the random approach. Each graph shows a distribution of betas for respective cardiac biometrics, with red lines showing *beta* estimated with the time-sensitive method.

For BPM, the random method produced reliability estimates that were very close to those resulting from the time-sensitive approach, if slightly lower during wakeful time, $M_{\text{beta\_sleep}}$ = 0.99, *SD* = 0.02, $M_{\text{beta\_wakeful}}$ = 0.89, *SD* = 0.058. For HRV during sleep, reliability from the random method was slightly lower than from the time-sensitive approach, $M_{\text{beta\_sleep}}$ = 0.98, *SD* = 0.03, supporting our assumptions. However, for wakeful HRV, the random approach resulted in somewhat higher estimated reliability than that produced by the time-sensitive method, $M_{\text{beta\_wakeful}}$ = 0.22, *SD* = 0.13.

To summarize, the two methods of estimating within-participant reliability revealed that both BPM and HRV were highly reliable during sleep, BPM was also very reliable during wakeful time, yet reliability of HRV during wakeful time was drastically lower. The two methods diverged in assessment of the latter, and not in the predicted direction, with the time-sensitive method yielding much lower estimates of reliability. Notice, however, that the number of datapoints obtained for HRV during wakeful hours was much lower than for sleep-time HRV or for BPM during wakefulness (see descriptive statistics above). Therefore, the amount of time separating successive datapoints must have been particularly long for wakeful HRV, likely exceeding the period during which we would expect such a volatile measure as HRV to be stable. The fact that the range of reliability estimates obtained with the random method was extremely wide (0.2–1) supports this reasoning.

4.3. Interim Discussion

We have demonstrated how between-person and within-person reliability can be estimated in data readily available from a commercial wearable sensor of cardiac biometrics. For the particular device tested here, between-person reliability as assessed with an ICC was excellent for sleep-time HR and HRV, but only moderate for wakeful biometrics. Within-person reliability, assessed using the split-half and the mixed model regression approach, was near-perfect for HR during sleep as well as wakeful time. However, HRV had high within-person reliability during sleep, but not during wakefulness.

It is worth noting that periods of lower reliability in this study coincided with periods in which fewer datapoints were obtained (lower measurement density). This could be taken to suggest that increased measurement density contributes to greater reliability. Yet, this coincidence in the present study should not be interpreted too strongly, since heart rate during wakeful times had a relatively high within-participant reliability even in the face of relatively fewer datapoints (lower measurement density). Further studies should investigate whether and how much measurement density contributes to higher reliability.

One immediate consequence of the compromised reliability of a measure is the reduced ability to detect its relationships with other variables [23]. We demonstrate this in what follows by testing whether BPM and HRV can be predicted from subjectively reported emotional states of the participants. Multiple laboratory studies showed that stress and cardiac biomarkers are strongly associated, and this relationship was recently replicated with wearable sensors [42,56]. We had no prior hypothesis as to which of the biomarkers (BPM, HRV) would produce a stronger association if they were measured with equal fidelity.

4.4. Correlations between Biomarkers and Subjective Emotions

To analyze subjective emotions captured with the EMA, we averaged the four negative emotions on each prompt (irritable, afraid, nervous, angry), and the two positive emotions (happy, energetic). We then averaged responses to all the prompts within one day, which resulted in two scores per day: one for positive and one for negative emotions.

We then used these two scores (negative and positive emotions) as predictors in mixed model regressions with participants as the random factor. We first tested wakeful BPM (and wakeful HRV in separate analyses) on the concurrent day as the dependent variables. Then, we tested the same models on sleep BPM (and sleep HRV in separate analyses), either on the preceding night (two models) or the following night (two more models). This meant that six models were tested in total, prompting us to use Bonferroni corrections to test for significance. The strongest relationship in all of these involved a negative relationship between sleep-BPM and lower negative mood reports on the subsequent reporting day, *beta* = -2.11, $t(40.25) = -2.745$, $p = 0.054$ (Bonferroni corrected). This meant that when a participant experienced higher sleep-BPM they were less likely to report negative emotions on the following day; when they experienced lower sleep-BMP, they were more likely to report negative emotions on the next day (see Figure 12). Wakeful-BPM was not reliably associated with either positive or negative emotions, *ps* > 0.5 (uncorrected).

For HRV, the strongest effect was also one where relatively higher sleep-HRV on a given night predicted greater negative emotions on the subsequent day, *beta* = 4.45, $t(40.5) = 1.79$, $p = 0.081$ (uncorrected). Wakeful-HRV was not associated with either positive or negative emotions, *ps* > 0.2 (uncorrected).

Positive emotions were not correlated with either of the biometrics tested, all *ps* > 0.13 (uncorrected).

4.5. Interim Discussion

Our attempt to test the association between mood during a day with cardiac biometrics concurrently (wakeful), on the preceding or the following night, revealed a predictive relationship between night-time BPM and mood on the subsequent day. We cannot tell from this study whether the difference between the results for BPM and HRV stems from a difference in reliability only, or for other reasons unrelated to measurement fidelity.

However, the difference between wakeful and sleep BPM in predicting daytime mood is surprising, given that both mood and cardiac biometrics change rapidly, so the closer in time that they are measured, the higher we would assume the association to be. It is therefore plausible that the association between daytime BPM and mood was compromised by the lower reliability of the daytime measurement.

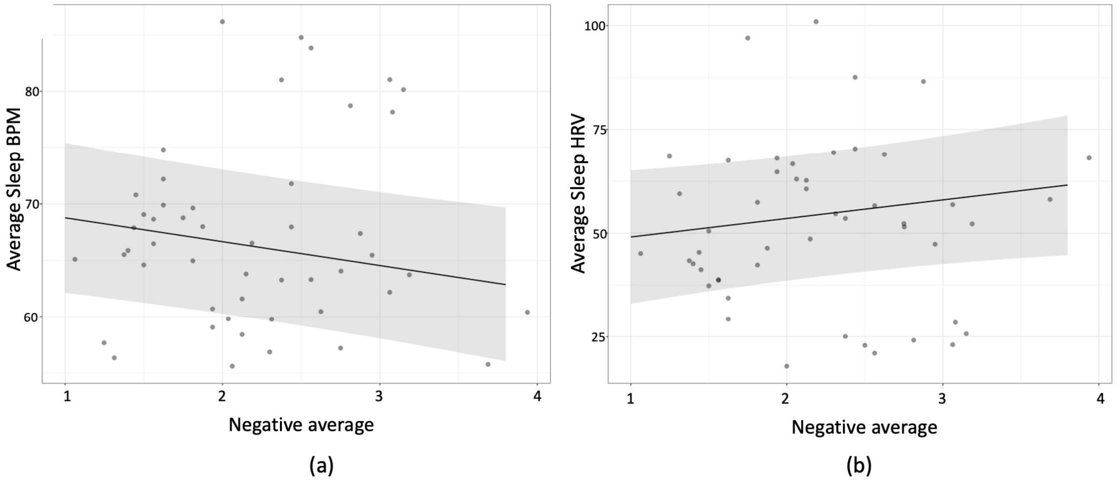

Figure 12. Association between negative emotions and sleep HR(V). Predicting sleep-BPM **panel** (**a**) and sleep-HRV **panel** (**b**) from negative emotions on a subsequent day. Grey area represents 95% CI.

5. General Discussion

The validity of data from wearable sensors is now thought to be quite good [1,3,4,6,7,9]. However, the reliability of the biometrics captured by these devices in daily life has so far been assumed to be high, but it has rarely been tested systematically. Interestingly, recent reports of using wearable sensors of HR and HRV outside of laboratory or clinical settings have revealed that validity of data from wearable sensors (i.e., correlations with a criterion, usually a medical-grade wearable device) in the conditions of everyday life is lower than expected from laboratory studies where participants are typically at quiet rest [57,58]. The finding by [19] that the reliability of a wearable (Empatica E4) HRV measurement across 24 h was unacceptably low offers an account of low validity of wearable sensors outside of a lab. Our examination of wearable data reliability in the present study also suggests that daytime measurements of HR and HRV are highly unreliable. Yet, rather than be discouraged by these data, we suggest focusing on *how* wearable sensors can be used to deliver data that is reliable. This can be achieved by assessing the sensor's reliability in different situations, as we performed here, for periods of sleep and wakefulness. This approach has been taken previously in the fields of movement science and athletics, where there was a growing awareness of the importance of testing the reliability of many forms of sensing technology in varying environmental and situational contexts [59–62].

Here we applied the theory of reliability developed for psychological questionnaires to physiological measurements obtained with a wearable device. In doing so, it is important to keep in mind the research goals and questions. If measurements are performed for comparisons between persons, between-participant reliability should be assessed, e.g., using ICC. If, however, the aim of the measurement is to detect different states within the same person, within-person reliability should be estimated, e.g., using a combination of the split-half and the mixed-model approaches. While estimating between-participant reliability is straightforward, estimating within-person reliability, without additional devices and measurements, must take into account the time-sensitive nature of the measurements being made. In the case of HR and HRV, the passing of time is a critical variable. Other physiological measurements will likely have similar considerations that are specific to the type of measurement being made.

We applied this approach to commercially available PPG sensors of cardiac biometrics, showing how between- and within-person reliability could be estimated from open-source data. The results showed that both between- and within-participant reliability of heart rate measurement varies for the different brands of wearables. Across different brands, it also varies for the different levels of the user's activity. This suggests that even the most reliable of the sensors tested (Apple Watch, Biovotion) may produce more or less reliable measurements in different circumstances.

Focusing on the Biostrap wearable sensor with our own data, we found that the between-participant reliability of HR and HRV was excellent during sleep (ICC > 0.75), but only fair during wakefulness (ICC [0.4 0.6]). Within-participant reliability of HRV was also found to be higher during sleep than during wakefulness. Finally, we found that correlations of HR and HRV with a second variable—in our case, subjectively reported mood—were stronger for the most reliable metric, sleep-time BPM. Taken as a whole, the present data suggest that the wearable sensor we used (Biostrap) provides data that are highly reliable during sleep, and less so during wakefulness.

The most popular testing of wearable devices has focused so far on measuring their validity during different levels of physical activity [1,53]. This is not surprising given that the early vision for the application of wearable devices was to regulate exercise load for performance optimization. This is still the primary use of many wearable devices today [53,63]. These studies have shown that that some measurements cannot be taken reliably during physical activity [63]. At the same time, other studies have shown that the application of wearables is not limited to detecting acute events (such as exercise load or acute stress), but that they can be useful in indexing slower fluctuations in the user's state, such as overall physical shape or allostatic stress, which affect one's long-term health and wellbeing [64,65]. This development opens up a unique research opportunity to measure psychophysiology longitudinally, across a variety of real-world contexts and extensive time periods [22]. Furthermore, this purpose might be best achieved with measurements taken during acute events or during recovery after those events. Indeed, a large body of studies have begun to investigate stress by focusing on recovery after stress, rather than on what is going on at the moment of acute stress [66]. The hope is that by systematically quantifying measurement reliability in different circumstances, researchers will eventually be able to make informed choices about specific wearable devices and measurement procedures that meet their research goals.

6. Conclusions

Wearable sensors offer a unique opportunity to measure physiology longitudinally and in various real-world contexts. This ecological benefit comes at a cost of a potentially compromised fidelity of the measurement. Here, we show that the reliability of a wearable sensor is not fixed but varies across different contexts and circumstances. Estimating reliability is thus a useful way to quantify measurement fidelity for a particular research question, a specific new experimental procedure, or a special target population.

Author Contributions: Conceptualization, V.D., G.D. and J.T.E.; methodology V.D. and J.T.E.; software V.D., C.Z. and O.B.; formal analysis, V.D. and J.T.E.; resources, G.D. and J.T.E.; writing—original draft preparation, V.D.; writing—review and editing, J.T.E., O.B. and G.D.; supervision, J.T.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Discovery Grant to JTE from the Natural Sciences and Engineering Council of Canada and by HealthQb Technologies, Inc. V.D.'s involvement in this research was funded by the Mitacs Accelerate Postdoctoral fellowship (IT23473).

Institutional Review Board Statement: The study procedures were approved by the Research Ethics Board of the University of British Columbia (approval number H19-01197).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data, materials, and the analysis code for this study are available at https://zenodo.org/badge/latestdoi/520639317 (accessed on 14 March 2023).

Conflicts of Interest: GD is a CEO of HealthQb Technologies, OB and CZ are employees of the same. HealthQb Technologies had no role in the design of the study, in the collection, analyses, or interpretation of data, or in the decision to publish the results.

References

1. Barrios, L.; Oldrati, P.; Santini, S.; Lutterotti, A. Evaluating the Accuracy of Heart Rate Sensors Based on Photoplethysmography for in-the-Wild Analysis. In Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, Trento, Italy, 20–23 May 2019; pp. 251–261.
2. Bent, B.; Goldstein, B.A.; Kibbe, W.A.; Dunn, J.P. Investigating Sources of Inaccuracy in Wearable Optical Heart Rate Sensors. *npj Digit. Med.* **2020**, *3*, 18. [CrossRef] [PubMed]
3. Hernando, D.; Roca, S.; Sancho, J.; Alesanco, Á.; Bailón, R. Validation of the Apple Watch for Heart Rate Variability Measurements during Relax and Mental Stress in Healthy Subjects. *Sensors* **2018**, *18*, 2619. [CrossRef] [PubMed]
4. Kinnunen, H.; Rantanen, A.; Kenttä, T.; Koskimäki, H. Feasible Assessment of Recovery and Cardiovascular Health: Accuracy of Nocturnal HR and HRV Assessed via Ring PPG in Comparison to Medical Grade ECG. *Physiol. Meas.* **2020**, *41*, 04NT01. [CrossRef] [PubMed]
5. Koskimäki, H.; Kinnunen, H.; Kurppa, T.; Röning, J. How Do We Sleep: A Case Study of Sleep Duration and Quality Using Data from Oura Ring. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, Singapore, 8–12 October 2018; pp. 714–717.
6. Menghini, L.; Gianfranchi, E.; Cellini, N.; Patron, E.; Tagliabue, M.; Sarlo, M. Stressing the Accuracy: Wrist-worn Wearable Sensor Validation over Different Conditions. *Psychophysiology* **2019**, *56*, e13441. [CrossRef]
7. Steinberg, B.A.; Yuceege, M.; Mutlu, M.; Korkmaz, M.H.; van Mourik, R.A.; Dur, O.; Chelu, M.; Marrouche, N. Utility of A Wristband Device as a Portable Screening Tool for Obstructive Sleep Apnea. *Circulation* **2017**, *136*, A19059.
8. van Lier, H.G.; Pieterse, M.E.; Garde, A.; Postel, M.G.; de Haan, H.A.; Vollenbroek-Hutten, M.M.R.; Schraagen, J.M.; Noordzij, M.L. A Standardized Validity Assessment Protocol for Physiological Signals from Wearable Technology: Methodological Underpinnings and an Application to the E4 Biosensor. *Behav. Res.* **2020**, *52*, 607–629. [CrossRef]
9. Dur, O.; Rhoades, C.; Ng, M.S.; Elsayed, R.; van Mourik, R.; Majmudar, M.D. Design Rationale and Performance Evaluation of the Wavelet Health Wristband: Benchtop Validation of a Wrist-Worn Physiological Signal Recorder. *JMIR Mhealth Uhealth* **2018**, *6*, e11040. [CrossRef]
10. Salehizadeh, S.; Dao, D.; Bolkhovsky, J.; Cho, C.; Mendelson, Y.; Chon, K. A Novel Time-Varying Spectral Filtering Algorithm for Reconstruction of Motion Artifact Corrupted Heart Rate Signals During Intense Physical Activities Using a Wearable Photoplethysmogram Sensor. *Sensors* **2015**, *16*, 10. [CrossRef]
11. Scully, C.G.; Lee, J.; Meyer, J.; Gorbach, A.M.; Granquist-Fraser, D.; Mendelson, Y.; Chon, K.H. Physiological Parameter Monitoring from Optical Recordings with a Mobile Phone. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 303–306. [CrossRef]
12. Han, D.; Bashar, S.K.; Lázaro, J.; Mohagheghian, F.; Peitzsch, A.; Nishita, N.; Ding, E.; Dickson, E.L.; DiMezza, D.; Scott, J.; et al. A Real-Time PPG Peak Detection Method for Accurate Determination of Heart Rate during Sinus Rhythm and Cardiac Arrhythmia. *Biosensors* **2022**, *12*, 82. [CrossRef]
13. Reiss, A.; Indlekofer, I.; Schmidt, P.; Van Laerhoven, K. Deep PPG: Large-Scale Heart Rate Estimation with Convolutional Neural Networks. *Sensors* **2019**, *19*, 3079. [CrossRef] [PubMed]
14. Zhang, Z.; Pi, Z.; Liu, B. TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-Type Photoplethysmographic Signals During Intensive Physical Exercise. *IEEE Trans. Biomed. Eng.* **2015**, *62*, 522–531. [CrossRef] [PubMed]
15. Tamura, T. Current Progress of Photoplethysmography and SPO_2 for Health Monitoring. *Biomed. Eng. Lett.* **2019**, *9*, 21–36. [CrossRef] [PubMed]
16. Biswas, D.; Simoes-Capela, N.; Van Hoof, C.; Van Helleputte, N. Heart Rate Estimation from Wrist-Worn Photoplethysmography: A Review. *IEEE Sensors J.* **2019**, *19*, 6560–6570. [CrossRef]
17. Berntson, G.G.; Quigley, K.S.; Lozano, D. *8 Cardiovascular Psychophysiology*; Cambridge University Press: Cambridg, UK, 2007; p. 29.
18. Zheng, Y.; Poon, C.C.Y. Wearable Devices and Their Applications in Surgical Robot Control and P-Medicine. In Proceedings of the 2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Nanchang, China, 4–6 May 2016; pp. 659–663.
19. Van Voorhees, E.E.; Dennis, P.A.; Watkins, L.L.; Patel, T.A.; Calhoun, P.S.; Dennis, M.F.; Beckham, J.C. Ambulatory Heart Rate Variability Monitoring: Comparisons Between the Empatica E4 Wristband and Holter Electrocardiogram. *Psychosom. Med.* **2022**, *84*, 210–214. [CrossRef] [PubMed]
20. Revelle, W.; Condon, D.M. Reliability from α to ω: A Tutorial. *Psychol. Assess.* **2019**, *31*, 1395–1411. [CrossRef]
21. Fuller, D.; Colwell, E.; Low, J.; Orychock, K.; Tobin, M.A.; Simango, B.; Buote, R.; Van Heerden, D.; Luan, H.; Cullen, K.; et al. Reliability and Validity of Commercially Available Wearable Devices for Measuring Steps, Energy Expenditure, and Heart Rate: Systematic Review. *JMIR Mhealth Uhealth* **2020**, *8*, e18694. [CrossRef]

22. Kleckner, I.R.; Feldman, M.J.; Goodwin, M.S.; Quigley, K.S. Framework for Selecting and Benchmarking Mobile Devices in Psychophysiological Research. *Behav. Res.* **2021**, *53*, 518–535. [CrossRef]
23. Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 72–101. [CrossRef]
24. Johnson, H.G. An Empirical Study of the Influence of Errors of Measurement upon Correlation. *Am. J. Psychol.* **1944**, *57*, 521–536. [CrossRef]
25. Nimon, K.; Zientek, L.R.; Henson, R.K. The Assumption of a Reliable Instrument and Other Pitfalls to Avoid When Considering the Reliability of Data. *Front. Psychol.* **2012**, *3*, 102. [CrossRef]
26. Sklar, A.Y.; Goldstein, A.; Hassin, R.R. Regression to the Mean Does Not Explain Away Nonconscious Processing: A Critical Review of Shanks 2017. *Exp. Psychol.* **2021**, *68*, 130–136. [CrossRef]
27. Fatisson, J.; Oswald, V.; Lalonde, F. Influence Diagram of Physiological and Environmental Factors Affecting Heart Rate Variability: An Extended Literature Overview. *Heart Int.* **2016**, *11*, heartint.500023. [CrossRef]
28. Shaffer, F.; Ginsberg, J.P. An Overview of Heart Rate Variability Metrics and Norms. *Front. Public Health* **2017**, *5*, 258. [CrossRef]
29. Hedge, C.; Powell, G.; Sumner, P. The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences. *Behav. Res.* **2018**, *50*, 1166–1186. [CrossRef]
30. Printz, M.P.; Jaworski, R.L. Hypertension: Overview. In *Encyclopedia of Endocrine Diseases*; Elsevier: Amsterdam, The Netherlands, 2004; pp. 369–380.
31. Sommerfeldt, S.L.; Schaefer, S.M.; Brauer, M.; Ryff, C.D.; Davidson, R.J. Individual Differences in the Association Between Subjective Stress and Heart Rate Are Related to Psychological and Physical Well-Being. *Psychol. Sci.* **2019**, *30*, 1016–1029. [CrossRef]
32. Zhang, D.; Shen, X.; Qi, X. Resting Heart Rate and All-Cause and Cardiovascular Mortality in the General Population: A Meta-Analysis. *Can. Med. Assoc. J.* **2016**, *188*, E53–E63. [CrossRef]
33. Laborde, S.; Mosley, E.; Thayer, J.F. Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research—Recommendations for Experiment Planning, Data Analysis, and Data Reporting. *Front. Psychol.* **2017**, *8*, 213. [CrossRef]
34. Natarajan, A.; Pantelopoulos, A.; Emir-Farinas, H.; Natarajan, P. Heart Rate Variability with Photoplethysmography in 8 Million Individuals: A Cross-Sectional Study. *Lancet Digit. Health* **2020**, *2*, e650–e657. [CrossRef]
35. Karjalainen, J.; Viitasalo, M. Fever and Cardiac Rhythm. *Arch. Intern. Med.* **1986**, *146*, 1169–1171. [CrossRef]
36. Williams, D.W.; Koenig, J.; Carnevali, L.; Sgoifo, A.; Jarczok, M.N.; Sternberg, E.M.; Thayer, J.F. Heart Rate Variability and Inflammation: A Meta-Analysis of Human Studies. *Brain Behav. Immun.* **2019**, *80*, 219–226. [CrossRef] [PubMed]
37. Chowdhury, M.R.; Madanu, R.; Abbod, M.F.; Fan, S.-Z.; Shieh, J.-S. Deep Learning via ECG and PPG Signals for Prediction of Depth of Anesthesia. *Biomed. Signal Process. Control* **2021**, *68*, 102663. [CrossRef]
38. Kasaeyan Naeini, E.; Subramanian, A.; Calderon, M.-D.; Zheng, K.; Dutt, N.; Liljeberg, P.; Salantera, S.; Nelson, A.M.; Rahmani, A.M. Pain Recognition with Electrocardiographic Features in Postoperative Patients: Method Validation Study. *J. Med. Internet Res.* **2021**, *23*, e25079. [CrossRef] [PubMed]
39. Koenig, J.; Jarczok, M.N.; Ellis, R.J.; Hillecke, T.K.; Thayer, J.F. Heart Rate Variability and Experimentally Induced Pain in Healthy Adults: A Systematic Review: HRV Nociceptive Stimulation Review. *Eur. J. Pain* **2014**, *18*, 301–314. [CrossRef]
40. Lim, H.; Kim, B.; Noh, G.-J.; Yoo, S. A Deep Neural Network-Based Pain Classifier Using a Photoplethysmography Signal. *Sensors* **2019**, *19*, 384. [CrossRef]
41. Brosschot, J.F.; Van Dijk, E.; Thayer, J.F. Daily Worry Is Related to Low Heart Rate Variability during Waking and the Subsequent Nocturnal Sleep Period. *Int. J. Psychophysiol.* **2007**, *63*, 39–47. [CrossRef]
42. Hovsepian, K.; al'Absi, M.; Ertin, E.; Kamarck, T.; Nakajima, M.; Kumar, S. CStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing—UbiComp '15, Osaka, Japan, 7–11 September 2015; pp. 493–504.
43. Perini, R.; Veicsteinas, A. Heart Rate Variability and Autonomic Activity at Rest and during Exercise in Various Physiological Conditions. *Eur. J. Appl. Physiol.* **2003**, *90*, 317–325. [CrossRef]
44. Tulppo, M.P.; Mäkikallio, T.H.; Seppänen, T.; Laukkanen, R.T.; Huikuri, H.V. Vagal Modulation of Heart Rate during Exercise: Effects of Age and Physical Fitness. *Am. J. Physiol. Heart Circ. Physiol.* **1998**, *274*, H424–H429. [CrossRef]
45. Buchheit, M.; Simon, C.; Charloux, A.; Doutreleau, S.; Piquard, F.; Brandenberger, G. Heart Rate Variability and Intensity of Habitual Physical Activity in Middle-Aged Persons. *Med. Sci. Sport. Exerc.* **2005**, *37*, 1530–1534. [CrossRef]
46. Rennie, K.L. Effects of Moderate and Vigorous Physical Activity on Heart Rate Variability in a British Study of Civil Servants. *Am. J. Epidemiol.* **2003**, *158*, 135–143. [CrossRef]
47. Janse van Rensburg, D.C.; Ker, J.A.; Grant, C.C.; Fletcher, L. Effect of Exercise on Cardiac Autonomic Function in Females with Rheumatoid Arthritis. *Clin. Rheumatol.* **2012**, *31*, 1155–1162. [CrossRef]
48. Routledge, F.S.; Campbell, T.S.; McFetridge-Durdle, J.A.; Bacon, S.L. Improvements in Heart Rate Variability with Exercise Therapy. *Can. J. Cardiol.* **2010**, *26*, 303–312. [CrossRef]
49. Bartko, J.J. The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychol. Rep.* **1966**, *19*, 3–11. [CrossRef] [PubMed]
50. Cicchetti, D.V. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychol. Assess.* **1994**, *6*, 284–290. [CrossRef]

51. Liljequist, D.; Elfving, B.; Skavberg Roaldsen, K. Intraclass Correlation—A Discussion and Demonstration of Basic Features. *PLoS ONE* **2019**, *14*, e0219854. [CrossRef]
52. Van Norman, E.R.; Parker, D.C. A Comparison of Split-Half and Multilevel Methods to Assesss the Reliability of Progress Monitoring Outcomes. *J. Psychoeduc. Assess.* **2018**, *36*, 616–627. [CrossRef]
53. Thomson, E.A.; Nuss, K.; Comstock, A.; Reinwald, S.; Blake, S.; Pimentel, R.E.; Tracy, B.L.; Li, K. Heart Rate Measures from the Apple Watch, Fitbit Charge HR 2, and Electrocardiogram across Different Exercise Intensities. *J. Sport. Sci.* **2019**, *37*, 1411–1419. [CrossRef]
54. Bent, B.; Dunn, J. BigIdeasLab_STEP: Heart Rate Measurements Captured by Smartwatches for Differing Skin Tones (Version 1.0). *PhysioNet* **2021**.
55. Jarchi, D.; Salvi, D.; Velardo, C.; Mahdi, A.; Tarassenko, L.; Clifton, D.A. Estimation of HRV and SpO2 from Wrist-Worn Commercial Sensors for Clinical Settings. In Proceedings of the 2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN), Las Vegas, NV, USA, 4–7 March 2018; pp. 144–147.
56. Coutts, L.V.; Plans, D.; Brown, A.W.; Collomosse, J. Deep Learning with Wearable Based Heart Rate Variability for Prediction of Mental and General Health. *J. Biomed. Inform.* **2020**, *112*, 103610. [CrossRef] [PubMed]
57. Galarnyk, M.; Quer, G.; McLaughlin, K.; Ariniello, L.; Steinhubl, S.R. Usability of a Wrist-Worn Smartwatch in a Direct-to-Participant Randomized Pragmatic Clinical Trial. *Digit. Biomark.* **2019**, *3*, 176–184. [CrossRef] [PubMed]
58. Sneddon, G.; Carlin, C. P129 User Experience and Accuracy of Continuous Cardio-Respiratory Physiology Data from a Wearable Photoplethysmography Wristband. *Thorax* **2019**, *74*, A160. [CrossRef]
59. Evenson, K.R.; Spade, C.L. Review of Validity and Reliability of Garmin Activity Trackers. *J. Meas. Phys. Behav.* **2020**, *3*, 170–185. [CrossRef]
60. Kobsar, D.; Charlton, J.M.; Tse, C.T.F.; Esculier, J.-F.; Graffos, A.; Krowchuk, N.M.; Thatcher, D.; Hunt, M.A. Validity and Reliability of Wearable Inertial Sensors in Healthy Adult Walking: A Systematic Review and Meta-Analysis. *J. NeuroEng. Rehabil.* **2020**, *17*, 62. [CrossRef]
61. Kooiman, T.J.M.; Dontje, M.L.; Sprenger, S.R.; Krijnen, W.P.; van der Schans, C.P.; de Groot, M. Reliability and Validity of Ten Consumer Activity Trackers. *BMC Sports Sci. Med. Rehabil.* **2015**, *7*, 24. [CrossRef] [PubMed]
62. Straiton, N.; Alharbi, M.; Bauman, A.; Neubeck, L.; Gullick, J.; Bhindi, R.; Gallagher, R. The Validity and Reliability of Consumer-Grade Activity Trackers in Older, Community-Dwelling Adults: A Systematic Review. *Maturitas* **2018**, *112*, 85–93. [CrossRef]
63. Almeida, M.; Bottino, A.; Ramos, P.; Araujo, C.G. Measuring Heart Rate During Exercise: From Artery Palpation to Monitors and Apps. *Int. J. Cardiovasc. Sci.* **2019**, *32*, 396–407. [CrossRef]
64. Chuang, C.-C.; Ye, J.-J.; Lin, W.-C.; Lee, K.-T.; Tai, Y.-T. Photoplethysmography Variability as an Alternative Approach to Obtain Heart Rate Variability Information in Chronic Pain Patient. *J. Clin. Monit. Comput.* **2015**, *29*, 801–806. [CrossRef]
65. Koskimäki, H.; Kinnunen, H.; Rönkä, S.; Smarr, B. Following the Heart: What Does Variation of Resting Heart Rate Tell about Us as Individuals and as a Population. In Proceedings of the Adjunct 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, London, UK, 9 September 2019; pp. 1178–1181.
66. Allen, A.P.; Kennedy, P.J.; Cryan, J.F.; Dinan, T.G.; Clarke, G. Biological and Psychological Markers of Stress in Humans: Focus on the Trier Social Stress Test. *Neurosci. Biobehav. Rev.* **2014**, *38*, 94–124. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.