Original Paper

Validation of Heart Rate– and Electrodermal Activity–Related Signals From Commercial-Grade Wearable Smartwatches and Rings in Experimentally Induced Stress: Laboratory-Based Study

Jiaer Guo[1], MSc; Courtney L Pollock[2,3], PhD; Rebecca M Todd[4,5], PhD; Anthony T Herdman[5,6], PhD; Noah D Silverberg[3,4,5], PhD

[1]Graduate Program of Neuroscience, Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, BC, Canada
[2]Department of Physical Therapy, University of British Columbia, Vancouver, BC, Canada
[3]Centre for Aging SMART, Rehabilitation Research Program, Vancouver Coastal Health Research Institute, Vancouver, BC, Canada
[4]Department of Psychology, University of British Columbia, Vancouver, BC, Canada
[5]Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, BC, Canada
[6]School of Audiology and Speech Sciences, University of British Columbia, Vancouver, BC, Canada

Corresponding Author:

Noah D Silverberg, PhD
Department of Psychology
University of British Columbia
2136 West Mall
Vancouver, BC V6T 1Z4
Canada
Phone: +1-604-822-2755
Email: noah.silverberg@ubc.ca

Abstract

Background: Commercial wearable devices such as smartwatches and smart rings have potential for remote monitoring of stress physiology in health research. These devices are commonly equipped with photoplethysmography sensors to measure heart rate (HR) and HR variability (HRV) and/or electrodermal activity (EDA) sensors to measure skin conductance responses, considered indicators of stress. They require further validation before being used in clinical research.

Objective: The study aims to evaluate the sensitivity of HR and HRV measurements obtained from Nowatch X, Empatica EmbracePlus, and Oura Ring and EDA measurements from Nowatch X, Empatica EmbracePlus, and Nuanic in controlled lab-based stress induction tasks, as well as the correspondence between the measurements from the wearable and lab devices.

Methods: Sixty-two healthy participants completed 4 stress induction tasks while wearing the abovementioned wearable devices, along with gold standard lab equipment (MindWare). HR, HRV, EDA, and EDA-derived metrics were calculated and compared against the lab recordings during the stress induction tasks and interleaved rest periods. We used repeated measures ANOVAs to assess within-device changes during rest and stress protocols. Linear mixed effects models, Bland-Altman plots, and Pearson correlation coefficients evaluated the strength of association and agreement between measurements obtained from the wearable devices and the laboratory device.

Results: Nowatch's recording setting precluded EDA readings. EDA readings from Empatica were deemed unreliable and therefore not analyzed. Laboratory stress induction tasks elicited subjective stress and physiological responses as measured by both wearable and lab devices. Nuanic's EDA showed a strong association with laboratory EDA measures at a signal level (r=0.78). Nowatch HR and mean duration of normal-to-normal interbeat intervals (MeanNN) closely matched MindWare electrocardiogram (ECG) measures (r=0.87 and r=0.85, respectively), whereas the SD of normal-to-normal intervals (SDNN) and root mean square of successive differences (RMSSD) showed weaker associations (r=0.55 and r=0.42, respectively). Empatica also demonstrated a strong association with MindWare (r=0.85), while Oura Ring's processed HR did not correspond well (r=0.68) with MindWare. HRV from both Empatica and Oura Ring had a weaker correspondence with MindWare ECG–derived HRV (RMSSD; r=0.36 and r=0.41, respectively). Bland-Altman analyses showed relatively small mean biases for most measures. Wearable-derived HR tended to be slightly lower than ECG-derived HR, whereas the direction of bias

varied across EDA and HRV measures. Wearable skin conductance response amplitude and HR measures showed narrower limits of agreement and lower mean absolute errors, whereas HRV metrics generally showed wider limits of agreement and larger mean absolute errors.

Conclusions: Overall, the wearable devices showed good sensitivity to experimentally induced stress and strong correspondence with laboratory-based HR and EDA measures in a controlled environment, whereas HRV indices were less consistent. Further validation under ambulatory and free-living conditions is needed to determine the performance of these devices in real-world monitoring applications.

JMIR Biomed Eng 2026;11:e101911; doi: 10.2196/101911

Keywords: EDA; skin conductance; ECG; wearable device; heart rate variability; Empatica; Oura Ring; electrocardiogram; electrodermal activity

Introduction

Wearable sensor devices such as smartwatches and smart rings have become increasingly popular research tools for tracking human physiology in everyday life. Many commercially available devices are equipped with photoplethysmography (PPG) sensors to measure blood volume changes, from which heart rate (HR) and HR variability (HRV) can be derived [1,2]. Some devices are also fitted with dry electrodermal activity (EDA) sensors to measure skin conductance responses (SCRs), considered an index of sympathetic nervous system activation [3]. Data collected from the devices are stored on the research participant's smartphone or on a cloud service. Measuring HR, HRV, and EDA outside the laboratory while participants are engaging in their real-world daily activities is very appealing. However, few wearable sensor devices have been validated against laboratory-based gold standard measures, such as electrocardiogram (ECG) for HR and HRV and palmar EDA for SCRs [4-6]. It is essential to confirm the accuracy of wearable sensor devices before deploying them in health research.

Past research focused on validating HR and HRV from common devices such as Garmin or Apple Watch and Oura Ring, especially during sleep and exercise. For example, in Miller et al [7], measurement of HR and HRV from 6 wearable devices during sleep were validated against laboratory recordings. Among the 6 devices, Oura Ring (Generation 2) had the highest agreement for both HR and HRV during sleep. They also found that HR and HRV sampled during sleep (vs wakefulness) generally performed better, likely due to motion artifacts affecting the quality of the signal.

Very few studies have tested the validity of EDA sensors in wearable devices. The Empatica watch is appealing because it is equipped not only with PPG but also EDA sensors. However, studies reported a modest to low reliability and inconsistent accuracy in its raw signal [8] and EDA time and frequency parameters [9] when tested against lab devices. Nuanic (previously branded as Moodmetric) is a wearable ring designed to specifically measure EDA. It is less studied than wrist devices but had shown more promising results and greater similarity in EDA phasic components and other features compared to lab sensors [10,11].

These inconsistencies across devices underscore the importance of validating each wearable device and using a standardized validation protocol to evaluate different parameters. Van Lier et al [12] proposed a comprehensive approach to validating wearable physiological devices on three levels: (1) at the signal level, to assess agreement in raw signals from wearable devices and lab-based gold standard equipment; (2) at the parameter level, to evaluate signal-derived metrics such as number of SCRs and indices of HRV; and (3) at the event level, to assess the sensitivity of the device to distinguish between events or conditions known to differentially trigger sympathetic and parasympathetic system responses. Achieving high agreement at the signal level is likely challenging due to sensor quality, artifacts, and differences in sensor placement. Furthermore, from a practical perspective, parameter- and event-level measurements (HRV indices, detecting a stressful time period) may be more likely than the raw signal (eg, EDA trace and PPG readings) to be used as dependent variables in research. Evaluating the validity of these outcomes would require protocols that reliably evoke measurable physiological responses to generate variability at the parameter and event levels. With respect to event-level responses, experimentally induced stress protocols may be especially helpful for studying the validity of wearable sensor devices against laboratory equipment for measuring stress physiology, including HR, HRV, and EDA [13]. Previous studies evaluating wearable sensor device performance have used a variety of cognitive, social, and environmental stressor tasks [12,14].

The present study aimed to validate HR and HRV metrics from Nowatch X, Empatica EmbracePlus, and Oura Ring Gen 4 and EDA metrics from Nuanic, Nowatch X, and Empatica EmbracePlus against laboratory-based ECG and EDA recordings. We used multiple experimentally induced stressors, counterbalanced and interleaved with resting periods. Our goals were twofold: (1) to assess whether recordings from each device are sensitive to experimentally induced stress and (2) to evaluate the correspondence between wearable-derived and lab-based measurements both during rest sessions and stress induction sessions.

Methods

Participants

We recruited participants from the University of British Columbia Human Subject Pool, where undergraduate students taking psychology classes can earn extra course credit by participating in research. Informed e-consent was obtained before the study session started. Participants were informed that they would be participating in a device validation study while completing tasks that would elicit transient stress. Potential participants were excluded if they had any neurological or cardiac diagnosis and if they used medications known to affect the autonomic nervous system or cardiovascular system (eg, opioids, beta-blockers, and topiramate).

Ethical Considerations

The study protocol was approved by the Behavioral Research Ethics Board of the University of British Columbia (H24-02522). All participants received information about the study and gave written informed consent. Participation was voluntary, and participants were free to withdraw from the study at any time without penalty. Participants received course credits for their time in the study.

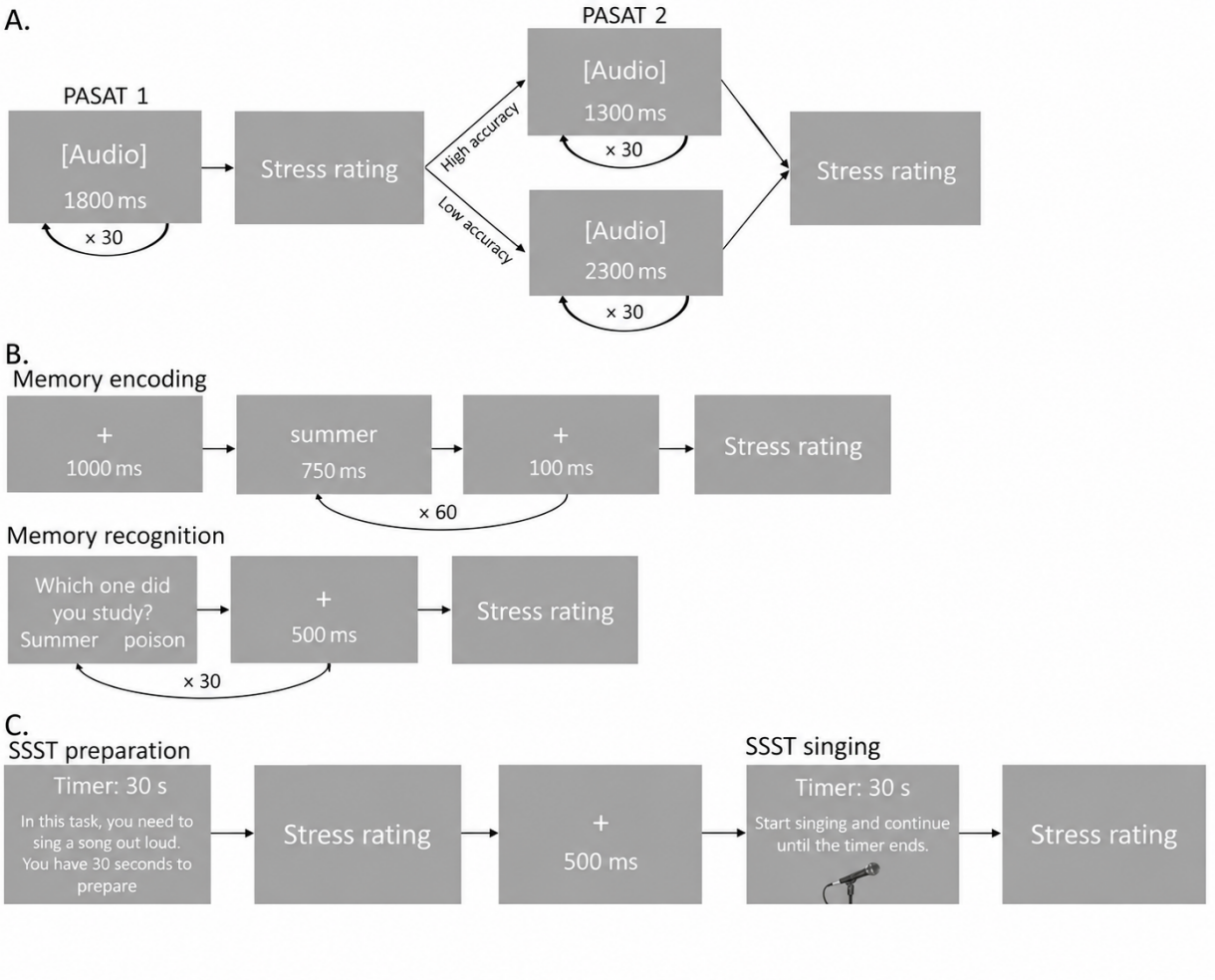
Devices

Two wearable watches, the Nowatch X (Polsen Health, formerly NOWATCH) and the Empatica EmbracePlus (Empatica Inc), and 2 wearable rings, the Nuanic ring (Nuanic Ltd) and the Oura Ring (Generation 4; Oura Health Ltd), were placed on participants, with hand and finger counterbalanced. At the time of the study, Nuanic was available in 4 sizes (7, 9, 10, and 12), and the rings were fitted onto the best-fitting finger if the sizes did not fit onto the counterbalanced finger. The Nuanic ring was placed on the same hand as the Oura Ring. Nowatch and Nuanic are relatively new to the market and are not yet widely adopted in research. In contrast, the Empatica watch and Oura Ring have been extensively used in validation and psychophysiological studies, and they offer access to selected data through their API with membership fees. For use as the gold standard reference, EDA and ECG data were collected using the MindWare Impedance Cardiograph & GSC 2 Amplifier (MindWare Technologies).

Stress Induction Tasks

Figure 1 shows a schematic depicting the stress induction tasks.

Figure 1. Task schematic of the stress induction tasks. (A) Paced Auditory Serial Addition Task (PASAT). (B) word recognition memory task. (C) Sing-a-Song Stress Test (SSST). Tasks were separated by a 3-minute resting phase and appeared in a pseudorandom order for each participant.



Paced Auditory Serial Addition Test

The Paced Auditory Serial Addition Test (PASAT) is a measure of cognitive processing speed and functions related to attention [15]. Prerecorded audios of single digits were presented from a computer at 1800-millisecond intervals, and participants were instructed to add each new digit to the one immediately before it and say the sum aloud. The research assistant sat across a desk, recording their responses. The initial interstimulus interval was kept shorter than that in the traditional PASAT in order to maximally elicit stress [14,16]. After the first block of 30 trials, a second block was administered, with the level of difficulty titrated to the participant's performance on the first block. The interstimulus interval of the audio was adjusted to 1300 milliseconds if the previous trial was not challenging enough (accuracy >86%); 2300 milliseconds if it was too challenging (accuracy <52%), risking participant disengagement; or kept at 1800 milliseconds if the first block was challenging enough (accuracy 52%-86%).

Sing-a-Song Stress Test

The Sing-a-Song Stress Test (SSST) was developed as an acute social stressor [17]. Participants were first instructed to prepare to sing a song of their choice aloud. This preparation phase lasted 30 seconds. Participants were then instructed to start singing aloud for 30 seconds. A 30-second timer was presented on the screen during both the preparation and the singing phases. Participants were encouraged to sing for the full 30 seconds, and if the participant did not start singing within the first 3 to 5 seconds, the research assistant gave some quick song suggestions (eg, "you can sing happy birthday or the ABC song"). The research assistant also encouraged the participant to continue singing if they stopped before the 30-second timer ended (eg, "keep going until the timer ends").

Word Recognition Memory Task

In the encoding phase, participants were instructed to study 60 words for a memory test, using a word list from Carpenter et al [18]. Words appeared on the screen one at a time for 750 milliseconds each, with 100-millisecond interstimulus intervals. This rapid presentation rate was intended to make the task stressful. To further elicit stress, participants were instructed just prior to the encoding phase that there was a 50% chance they would receive an electric shock from the electrodes attached to their hand (used for EDA collection) if they made a recognition error; however, no shock was to be delivered. The encoding phase was immediately followed by a recognition phase, during which a previously studied word and a new (foil) word appeared on the screen side by side, and participants used the left or right arrow key to select the studied word. We expected that recognition performance would be high (accuracy>90%) and that participants would become effectively unblinded to the threat of shock deception after making a few errors. We restricted analyses to the first one-third of the recognition phase for each participant, where multiple errors were less likely to have occurred yet.

Open Affective Standardized Image Set Image Viewing Task

An image viewing task was included in the study, where participants passively viewed 30 images from the Open Affective Standardized Image Set (OASIS) [19]. During data collection, the task did not elicit significant EDA responses and is therefore not discussed in this paper. More information about the task is included in Multimedia Appendix 1.

Procedures

It was recommended by the manufacturers (personal communication) of Nowatch and Empatica that sensors be warmed up before data collection session, so a research assistant wore the watches 15 minutes before each participant's testing session. Upon arrival, participants' ring sizes were measured, and they were asked to complete study-specific demographic and health history questions as REDCap surveys.

After the questionnaires were completed, participants were instructed to wash their hands with soap and water before electrode placement and then asked to wear the watch and 2 rings on the assigned hand and fingers. Two wet EDA electrodes of the lab equiment were placed on participant's nondominant palm. Three-lead wet ECG electrodes were placed on the right collar bone and on each side of the bottom ribs [20]. The skin areas were cleaned with alcohol wipes before placing the electrodes, and electrode wires were secured with tapes to minimize movement artifacts.

After the setup, the research assistant instructed the participant to sit comfortably in front of the computer to complete the 4 stress induction tasks (PASAT, word recognition memory task, OASIS, and SSST) in a pseudorandom order while resting their nondominant hand on the desk (Figure 2). Because Oura rings did not provide continuous measurement of daytime HR and HRV at a resolution higher than every 5 minutes when in their default setting, we turned on the "unguided session" feature which requires the testing hand to be motionless. When "excessive" motion is detected (exceeding a proprietary threshold), HR and HRV output are not generated. Therefore, we asked the participants to try to keep their nondominant hand as still as possible to ensure recording. Two unguided sessions were recorded (about 20 minutes each), and participants were given a break to rest their hands before a new recording started. The break always happened after a stress induction task ended.

Figure 2. Experimental setup on the nondominant hand. Illustration of the physiological recording setup showing the placement of devices on the participant's nondominant hand. The Nowatch was worn on the wrist, and the Nuanic ring and Oura Ring were placed on the fingers. Electrodermal activity electrodes were attached to the palm, with surgical tape securing the electrodes and wires. The hand rested on a towel on the desk during the experiment.



At the beginning of the experiment, the research assistant sat in the same room as the participant, off to their side (90°), to serve as a social stressor. Participants were told that the research assistant was there to monitor the sensor signals. Participants were also told to minimize interaction with the research assistant except when task instructions were given before each stress induction task started. The research assistant exited the testing room before the start of each rest session and then returned to the room before the next task began.

The 4 tasks were presented pseudorandomly using PsychoPy [21]. Each task lasted for approximately 2 minutes (except OASIS, which lasted about 11 minutes) and was preceded by a 3-minute resting session where the participants were instructed to close their eyes and listen to an audio clip from a movie designed to be used during resting-state functional magnetic resonance imaging [22]. At the beginning and end of each resting session and after each task block, participants were instructed to self-report their subjective stress levels using a slider on a visual analog scale from 0 (low stress) to 10 (high stress). A final resting session and stress rating were also implemented at the end of the fourth task. After the tasks were completed, the devices were taken off the participants, and the participants were debriefed about the study and the deception component.

Analysis

Extraction of Physiological Metrics

Event markers from PsychoPy containing absolute time stamps were recorded by MindWare at the beginning of the study, at each task's start and end, and at the end of the study. Each wearable device also had absolute time stamps at each data point. Event markers were merged with wearable datasets by matching the nearest time stamp for task segmentation. Physiological signals and extracted metrics from each device are summarized in Table 1. HR and 3 HRV metrics (mean duration of normal-to-normal interbeat intervals [MeanNN], SD of normal-to-normal intervals [SDNN], and root mean square of successive differences [RMSSD]) were analyzed from Nowatch. HR and RMSSD were analyzed from Oura Ring and Empatica. Two EDA metrics (number of SCRs [nSCR] and the average amplitude of SCRs [Amp_SCR]) and the cleaned EDA signal were analyzed from Nuanic.

Table 1. Physiological metrics and their extraction from each device.

Device and metric	Definition	Calculation/extraction
Nuanic		
EDA	Electrodermal activity	EDA signal throughout each rest and task phase
nSCR	Number of SCRs[a]	Total count of SCR detected in each rest and task phase
Amp_SCR	Mean amplitude of SCRs	Average amplitude of SCRs detected in each rest and task phase
Nowatch		
HR	Heart rate (beats per minute)	Average heart rate in each rest and task phase
MeanNN	Mean duration of normal-to-normal IBIs[b]	Average IBI in each rest and task phase
SDNN	Variability of normal-to-normal IBIs	SD of IBI in each rest and task phase
RMSSD	Root mean square of successive differences between adjacent normal-to-normal intervals	RMSSD of IBI in each rest and task phase

Device and metric	Definition	Calculation/extraction
Oura		
HR	Heart rate (beats per minute)	HR calculated by Oura Ring's proprietary algorithm every 5 seconds throughout the unguided session
RMSSD	Root mean square of successive differences between adjacent normal-to-normal intervals	HRV[c] calculated by Oura Ring's proprietary algorithm every 5 seconds throughout the unguided session
Empatica		
HR	Heart rate (beats per minute)	HR calculated by Empatica's proprietary algorithm every 1 minute
RMSSD	Root mean square of successive differences between adjacent normal-to-normal intervals	HRV calculated by Empatica's proprietary algorithm every 1 minute

[a]SCR: skin conductance response.
[b]IBI: interbeat interval.
[c]HRV: heart rate variability.

Data Cleaning and Processing

Task Behavior

One participant withdrew from the image viewing task and the end-of-study resting phase. A programming issue for the SSST resulted in missing ratings from 13 participants in the resting phase before the SSST and after the SSST preparation phase. The missing resting phase rating was interpolated as the average of the participants' 3 other resting phase ratings, and the same rating from the participant's singing phase was used as their rating for the preparation phase of the SSST.

EDA

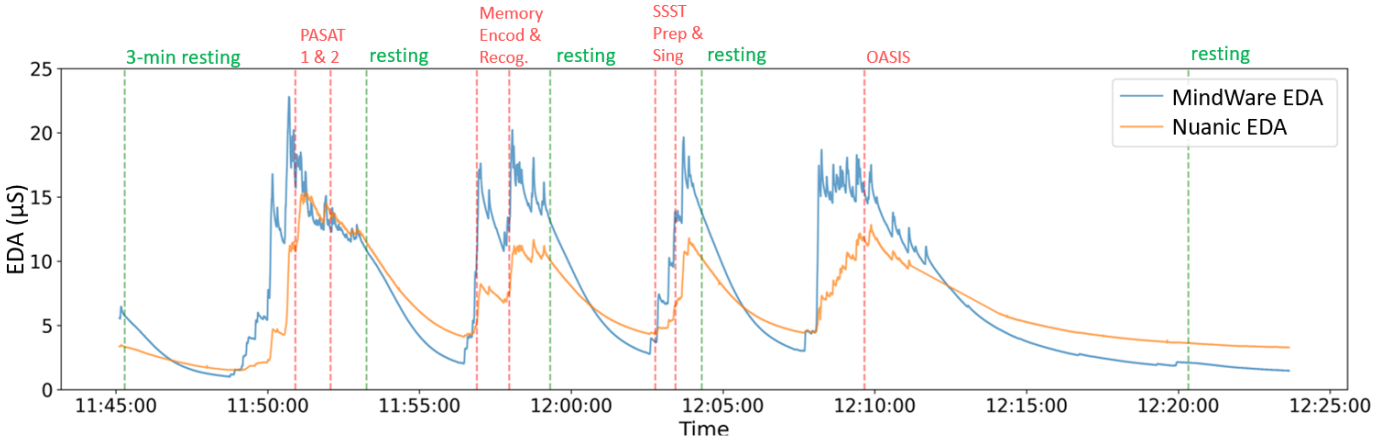
MindWare

MindWare EDA data were recorded at 500 Hz. An offline fourth-order low-pass filter of 10 Hz was applied to remove artifacts, and the cleaned EDA data were decomposed into phasic and tonic components using the MATLAB-based toolbox Ledalab (version 3.4.9) and its continuous decomposition analysis method [23]. The nSCR and Amp_SCR data were extracted. Data were downsampled to 50 Hz to optimize processing speed. EDA slopes greater than –10 to 10 μS/s were considered artifacts and removed from the data [24]. A response qualified as an SCR when the amplitude was greater than 0.05 μS.

Nuanic

EDA raw data were collected at 16 Hz. The majority of datasets were misaligned due to time referencing issues in the app. All available data were resynchronized by applying a constant time shift determined through visual inspection to align data morphologically (mean time shift –98, SD 91 seconds, range –367 to 0 seconds). The investigator performing the time correction was blinded to task onset and offset (ie, event markers were not visible). To determine the time shift, the investigator identified a distinctive EDA feature (eg, the first prominent peak following the resting period) that was clearly identifiable in both recordings. The time difference between the occurrence of the corresponding feature was recorded and used to determine the time shift. All calculated time shifts were applied in batch processing after all offsets had been determined. A representative time series is presented in Figure 3. Each participant's data were further segmented into different task phases based on absolute time stamps and event marks automatically labeled in MindWare, and the resynchronization ensured that the correct segment were extracted for each task period. Missing data points were observed due to (1) uneven time stamps created by the wearable recording system and (2) memory that was overwritten when data were not automatically synchronized from the device. Missing data segments shorter than 0.5 seconds were linearly interpolated using neighboring valid EDA samples immediately preceding and following each missing segment, and interpolation was performed within an extended window surrounding the gap. This step corrected all missing data points resulting from uneven time stamps. Larger segments that had less than 3 seconds of missing data were also linearly interpolated because EDA decomposition requires continuous data. Those interpolated larger segments were later excluded from analysis after decomposition (ie, when comparing at signal level with MindWare EDA). Segmented data with missing data segments larger than 3 seconds that were not at the start or the end of the recording were excluded from decomposition and analysis. A total of 46 of 574 (8.01%) segments were excluded. EDA features (nSCR and Amp_SCR) were extracted using the same method as MindWare EDA, except that no downsampling was applied.

Figure 3. Comparison of example electrodermal activity (EDA) recordings from MindWare and Nuanic devices across a 42-minute study session. The green dash line indicates when a resting session started. Each resting session lasted 3 minutes. The red line indicates when a stress task component started. Memory Encod & Recog: memory task encoding and recognition; OASIS: Open Affective Standardized Image Set; PASAT: Paced Auditory Serial Addition Task; SSST Prep & Sing: Sing-a-Song Stress Test preparation and singing.



Nowatch

Nowatch X contains a bioimpedance EDA sensor located at the ventral wrist position that measures electrical conductance between electrodes and skin. We discovered after designing the study that raw PPG and EDA readings cannot be collected simultaneously. This could be a system error in the collection software (NOWATCH iPad App version 2.1.8), which we reported to the manufacturer. We prioritized raw PPG data collection, and therefore EDA data were not collected as planned for the study.

Empatica

EDA data were recorded at 4 Hz at the ventral wrist. On data screening, we observed 4 participants with missing data, and data recorded from 53 of 61 (86.89%) participants showed a maximum amplitude of less than 2 µS and on average showed low variations (SD 0.17 µS). EDA data recorded from the wrist typically exhibit lower amplitudes and reduced variability compared to those obtained from the palms [25]; however, such consistently low amplitudes and limited variability raised suspicion that the data were recorded under suboptimal conditions (eg, watches not worn tightly enough) or that the sensors were not functioning properly. Even though we have alerted the manufacturer about data quality, an inspection showed the sensor to be functioning properly. Given the concerns about data quality and the uncertainty regarding the interpretability of these signals, these data were excluded from the present analyses.

ECG

MindWare

ECG data were recorded at 500 Hz. The Python-based toolbox Neurokit2 [26] was used to clean and extract HR and HRV metrics. HR and HRV were computed separately for each rest and task phase to enable comparison with Nowatch. Additionally, RMSSD was calculated in 25-second intervals to facilitate comparison with Oura-derived HRV.

Nowatch

We did not analyze the raw PPG signals, even though they were available, due to challenges in motion artifacts affecting PPG signal qualities. Rather, we analyzed the processed interbeat interval (IBI) and HR values calculated by Nowatch's proprietary algorithm at approximately 1 Hz and 0.1 Hz, respectively. IBIs longer than 1500 milliseconds or shorter than 300 milliseconds or values outside ±3 SDs were considered outliers, and those data points were excluded during offline data processing. Missing data shorter than 5 seconds were linearly interpolated. Data were segmented into different task phases by time stamps and event markers. In total, 44 of 706 (6.23%) segments were excluded due to excessive missing data. IBI data were further processed to calculate MeanNN, SDNN, and RMSSD. HR data were averaged by task phases, and 5 of 706 (0.71%) segments were excluded due to missing data. It is worth noting that Nowatch and Empatica allow for measuring and exporting raw PPG and 3-axis accelerometer signals. We chose to analyze the provided HR and IBI or HRV measurements instead to reduce data processing and validate Nowatch's and Empatica's proprietary algorithm that extracts HR and IBI or HRV.

Empatica

Similar to Nowatch, we did not analyze the raw PPG signals and instead analyzed the processed HR and RMSSD which were in 1-minute intervals. Problematic data were categorized into "worn_with_low_signal_quality," "worn_during_motion," "device_not_worn_correctly" automatically by Empatica's proprietary algorithm. On average, participants were missing 0.08% (SD 0.44%; max detected=100%, min detected=97.30%) of their HR data and 52.73% (SD 29.46%; max detected=97.14%, min detected=0%) of their HRV data. Four participants had no available HRV data and were not included in the analysis. Empatica data were not segmented by task; instead, MindWare ECG–derived HR and IBI were averaged to 1-minute intervals, matching the resolution of the Empatica data, and then compared.

Oura Ring

HR and HRV readings were missing if excessive movement was detected by the ring. Calculation of HRV (calculated as RMSSD by Oura Ring's proprietary algorithm) required more sedentary time periods and valid data than those required for HR [27] and was more likely to result in unavailable data. On average, participants were missing 10.0% (SD 12.62%; max=69.70%, min=1.33%) of their HR data and 26.1% (SD 23.12%; max=100%, min=4.8%) of HRV data. HRV recordings in the first 30 seconds in each session (about 2.50%) were always missing due to the time needed to calculate HRV. To compare the Oura Ring reading with MindWare's ECG recordings, ECG-derived HR and IBI values were averaged using a 25-second window for every 5 seconds to match the resolution with that of the Oura Ring. Other window sizes were also examined but yielded worse results (Figure S1 in Multimedia Appendix 1).

Event-Level Device Responsiveness

Repeated measures ANOVAs were used here to assess the within-device or metric changes during the rest and stress protocols. Subjective stress ratings, EDA, and HR-related metrics (HR and HRV) from MindWare, Nuanic, and Nowatch were summarized to assess the levels of perceived and objective stress responses in each phase of the study tasks. Empatica and Oura Ring data were not summarized because their resolution was too low to be segmented. Repeated measures ANOVAs were conducted using task phase as the within-subject factor, followed by post hoc pairwise comparisons to identify significant differences between phases, specifically between the tasks and the immediately preceding rest period. Significance was considered at $P<.05$ for ANOVA and at $P<.05$ after false discovery rate correction for directional (1-tailed) post hoc tests. The 1-tailed directional tests were selected because we presumed that the stress level should elevate during the stress tasks compared to the rest periods. This approach also allowed us to evaluate the ability of each device to distinguish between rest and stress induction conditions.

Signal-Level Multilevel Analysis

Multilevel analyses were performed in R software (version 4.4.2; R Foundation for Statistical Computing) using the *lme4* and *lmerTest* packages. The purpose of this analysis was to evaluate the strength of association between physiological measurements (eg, EDA, nSCR, and HR) obtained from wearable devices and those from the laboratory gold standard system. We adapted a linear mixed effects model (LMM) from van der Mee et al [14] using gold standard measurements as the dependent variables, measurements from the wearable devices as the predictors (fixed effect), and subject-specific random intercepts and slopes (random effect) to account for individual differences in baseline levels and device performances. Within-subject correlations between devices were also calculated using (1) cross-correlation if the metrics were captured as time series data (ie, Nuanic's EDA, Empatica's and Oura Ring's HR and RMSSD) or (2) Pearson r if the metrics were derived from time series data (ie, Nuanic's nSCR and Amp_SCR and Nowatch's HR, MeanNN, SDNN, and RMSSD) and averaged across participants.

In addition to mixed effects modeling, agreements between wearable and MindWare measurements were assessed using Bland-Altman analysis, mean absolute error (MAE), and intraclass correlation coefficient (ICC). For each metric, Bland-Altman plots were constructed by plotting the difference between the wearable and MindWare measurements against their mean. The mean difference (bias) and the 95% limits of agreement (bias ± 1.96 SD) were calculated to evaluate systematic measurement bias and the extent of agreement between methods. MAE is calculated as the average absolute difference between the wearable and MindWare measurements. Agreement was also assessed using a 2-way random-effects, single-measure, absolute agreement ICC ($ICC_{2,1}$).

Results

Participant Characteristics

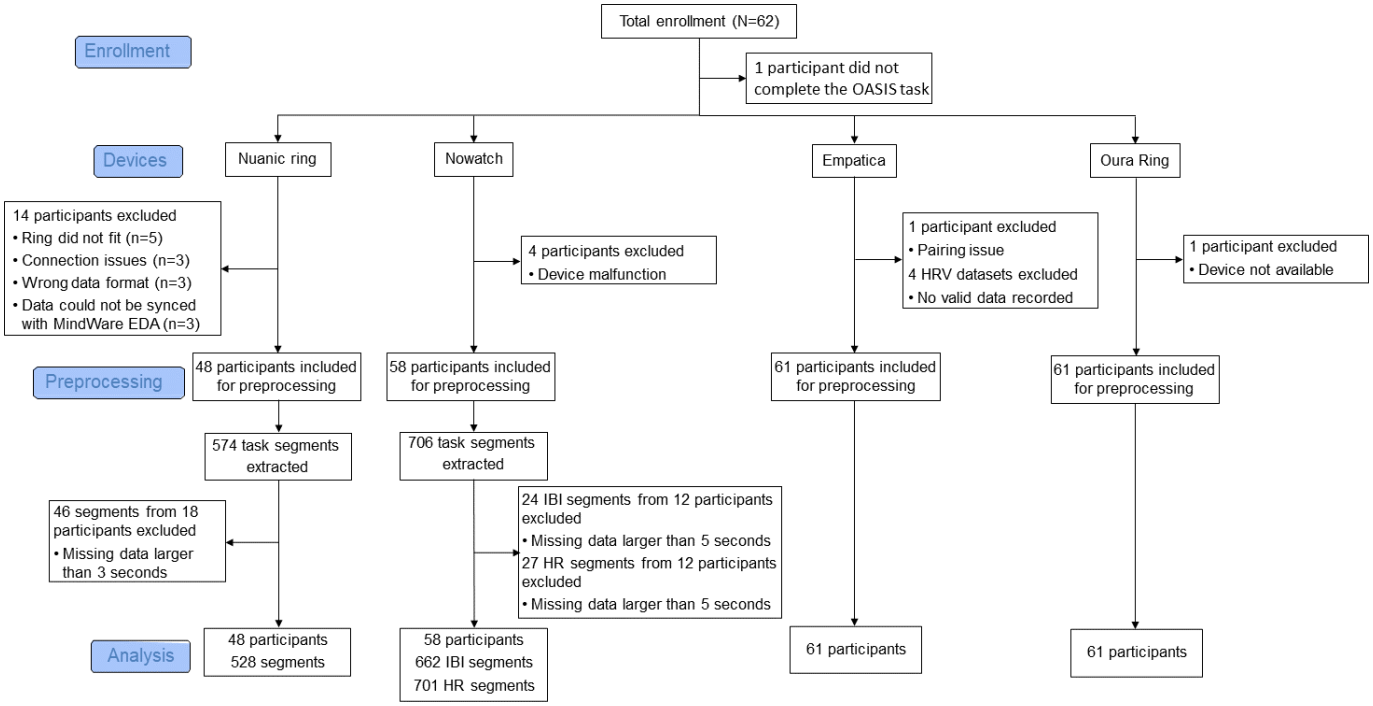
A total of 62 participants (52 women; mean age 20.60, SD 2.15 years) were recruited. Table 2 shows the sample demographics. Not all data from all devices for all participants were available for analysis. Reasons for missing and excluded data are detailed in Figure 4.

Table 2. Participant demographics (N=62).

Characteristic	n	%
Sex		
Female	52	83.87
Male	10	16.13
Gender		
Woman	50	80.65
Man	11	17.74
Nonbinary/third gender	1	1.61
Race		
East or Southeast Asian	31	50.00

Characteristic	n	%
White	22	35.48
South Asian	8	12.90
Black	4	6.45
Middle Eastern	3	4.84
Indigenous	2	3.23
Latinx	1	1.61
Race not listed above	2	3.23

Figure 4. Participant flow diagram showing participants flow and data inclusion. The final analysis included 48 participants (528 task segments) for Nuanic ring, 58 participants (662 IBI segments and 701 HR segments) for Nowatch, and 61 participants for both Empatica and Oura Ring. EDA: electrodermal activity; HR: heart rate; HRV: heart rate variability; IBI: interbeat interval; OASIS: Open Affective Standardized Image Set.



Device Responsiveness to Experimentally Induced Stress

Mean and SD of the subjective stress ratings and EDA recordings from each device are shown in Table 3. Statistically significant increases on EDA metrics during task phases relative to the immediately preceding rest periods are noted in Table 3, with full ANOVA results in Table S1 (Multimedia Appendix 1). Across task phases, MindWare EDA and ECG showed increased physiological measures relative to the immediately preceding rest periods, except for nSCR during memory recognition, consistent with heightened stress reactivity. Exceptions were also observed in HR-related indices (Table 4), where HR, MeanNN, and RMSSD did not show significant changes during the word recognition memory task. We found no statistical evidence for a change in SDNN during the initial PASAT or SSST phases, but there was a significant decrease in SDNN during the memory phases. However, no changes in RMSSD were observed during the memory phases.

Table 3. Average stress ratings and electrodermal activity (EDA) readings.

		EDA (µS), mean (SD)		nSCR[a], mean (SD)		Amp_SCR[b], mean (SD)	
	Stress score, mean (SD)	MindWare[c]	Nuanic[d]	MindWare	Nuanic	MindWare	Nuanic
Resting (PASAT[e])	2.1 (1.6)	6.0 (5.4)	5.8 (7.5)	3.0 (4.4)	4.6 (6.7)	0.3 (0.4)	0.2 (0.5)
PASAT_1	6.0 (1.9)[f]	9.9 (7.4)[f]	10.0 (12.2)[f]	9.1 (5.0)[f]	7.5 (3.8)[f]	0.7 (0.6)[f]	0.4 (0.3)[g]
PASAT_2	5.7 (1.9)[f]	9.6 (6.9)[f]	10.5 (12.9)[f]	9.0 (5.5)[f]	6.9 (4.2)[f]	0.5 (0.4)[f]	0.4 (0.4)[g]
Resting (Memory)	2.1 (1.5)	6.3 (5.5)	8.5 (13.6)	3.7 (5.4)	6.1 (10.8)	0.3 (0.4)	0.2 (0.2)
Memory Encoding	4.7 (1.8)[f]	8.1 (6.8)[f]	8.6 (11.7)[g]	5.0 (4.4)[f]	3.9 (3.5)[g]	0.6 (0.5)[f]	0.3 (0.3)[g]
Memory Recognition	3.9 (1.6)[f]	9.0 (6.7)[f]	9.0 (11.8)[g]	2.9 (2.0)[g]	2.1 (1.7)[g]	0.9 (0.9)[f]	0.3 (0.3)[g]

		EDA (μS), mean (SD)		nSCR[a], mean (SD)		Amp_SCR[b], mean (SD)	
	Stress score, mean (SD)	MindWare[c]	Nuanic[d]	MindWare	Nuanic	MindWare	Nuanic
Resting (SSST[h])	2.1 (1.4)	6.3 (5.2)	5.5 (4.2)	2.7 (4.4)	3.9 (5.7)	0.3 (0.4)	0.4 (1.1)
SSST Preparation	5.0 (2.0)[f]	9.4 (7.5)[g]	8.4 (8.0)[f]	5.7 (3.1)[f]	4.1 (2.6)[g]	1.2 (0.9)[f]	0.7 (0.7)[g]
SSST Singing	5.2 (2.1)[f]	11.7 (7.4)[f]	11.9 (11.6)[f]	5.7 (3.0)[f]	5.1 (2.9)[g]	1.0 (0.8)[f]	0.5 (0.5)[g]
Resting (End)	2.4 (1.8)	7.6 (5.3)	9.4 (11.9)	4.2 (5.3)	7.9 (10.4)	0.3 (0.4)	0.2 (0.2)

[a]nSCR: number of skin conductance responses.
[b]Amp_SCR: average amplitude of skin conductance responses.
[c]Gold standard laboratory device.
[d]Consumer-grade wearable device.
[e]PASAT: Paced Auditory Serial Addition Test.
[f]Significant increase from its prior resting session.
[g]No significant difference from its prior resting session.
[h]SSST: Sing-a-Song Stress Test.

Table 4. Heart rate (HR) and heart rate variability (HRV) indices.

	HR, mean (SD)		MeanNN[a], mean (SD)		SDNN[b], mean (SD)		RMSSD[c], mean (SD)	
	MindWare[d]	Nowatch[e]	MindWare	Nowatch	MindWare	Nowatch	MindWare	Nowatch
Resting (PASAT[f])	75.8 (11.8)	77.3 (11.4)	812.1 (135.9)	804.3 (127.8)	62.3 (29.9)	60.8 (28.0)	44.0 (29.5)	47.9 (24.3)
PASAT_1	83.8 (13.9)[g]	83.1 (14.0)[g]	736.7 (126.9)[g]	741.1 (122.2)[g]	54.5 (23.8)[h]	56.6 (25.7)[h]	32.8 (19.3)[g]	42.3 (24.3)[h]
PASAT_2	80.4 (12.8)[g]	79.9 (12.7)[g]	768.6 (129.5)[g]	767.0 (128.6)[g]	49.4 (24.0)[g]	50.7 (24.8)[g]	37.0 (26.2)[g]	43.8 (24.2)[h]
Resting (Memory)	75.7 (11.0)	76.8 (10.7)	810.4 (127.9)	805.9 (120.6)	63.5 (32.9)	62.5 (29.4)	44.1 (33.0)	48.2 (21.2)
Memory Encoding	77.2 (11.2)[h]	77.7 (10.9)[h]	795.3 (127.5)[h]	789.7 (114.0)[h]	49.5 (21.5)[g]	51.0 (25.7)[g]	39.2 (20.1)[h]	45.3 (20.1)[h]
Memory Recognition	78.4 (11.6)[g]	78.3 (11.1)[g]	783.0 (120.6)[g]	784.2 (119.1)[g]	42.1 (21.0)[g]	44.5 (25.6)[g]	38.5 (22.7)[h]	45.0 (28.3)[h]
Resting (SSST[i])	75.8 (11.0)	77.9 (10.6)	809.6 (124.9)	798.4 (118.2)	60.7 (31.5)	59.4 (29.2)	42.9 (30.4)	46.9 (22.0)
SSST Preparation	83.0 (11.8)[g]	81.1 (11.4)[g]	739.7 (117.1)[g]	755.4 (118.9)[g]	63.3 (31.5)[h]	56.8 (29.9)[h]	36.4 (28.3)[g]	40.0 (19.6)[h]
SSST Singing	90.5 (14.8)[g]	87.6 (14.3)[g]	682.4 (111.7)[g]	712.0 (140.5)[g]	56.9 (27.7)[h]	51.2 (28.8)[h]	30.1 (20.9)[g]	44.2 (36.0)[h]
Resting (End)	76.4 (11.1)	76.4 (10.6)	803.9 (125.3)	789.2 (153.1)	66.8 (34.9)	66.1 (28.8)	44.3 (30.7)	47.4 (19.8)

[a]MeanNN: mean duration of normal-to-normal interbeat intervals.
[b]SDNN: SD of normal-to-normal intervals.
[c]RMSSD: root mean square of successive differences.
[d]Gold standard laboratory device.
[e]Consumer-grade wearable device.
[f]PASAT: Paced Auditory Serial Addition Test.
[g]Significant increase from its prior resting session.
[h]No significant difference from its prior resting session.
[i]SSST: Sing-a-Song Stress Test.

Nuanic EDA was also sensitive to task manipulations, with especially prominent SCRs during the PASAT and SSST. In contrast, the word recognition memory task elicited little to no change. The average amplitude of SCRs was less sensitive than the number of SCRs to the stress induction task vs rest manipulation. All 3 EDA metrics were elevated during PASAT relative to the immediately preceding rest period. For other tasks, the number of SCRs increased, and the other SCR measures showed effects in the expected direction (higher during task than rest) but did not reach significance.

Average HR and HRV indices in each phase are described in Table 4. Nowatch-derived HR and MeanNN increased reliably during stress tasks and showed the same pattern of task sensitivity as MindWare ECG recordings. As with the lab device, SDNN from Nowatch showed minimal phase-related changes, and RMSSD did not respond consistently to any of the manipulations.

Overall, the experimental manipulation was effective in eliciting physiological stress responses and subjective perceived stress ratings, with the strongest effects observed during the PASAT and SSST. The word recognition memory task produced minimal physiological changes across measures, even with the threat of aversive stimulation.

Correspondence Between Wearable and Lab-Based Devices

The LMM results and correlations between the lab recordings and wearable devices across both task and resting conditions are reported in Table 5. The LMM revealed a strong association between Nuanic and lab-recorded EDA. Of the variance in lab EDA, 99.3% was explained by the model, which included Nuanic measurements and participant-level variation. On average, Nuanic EDA was highly correlated with lab EDA (r=0.78). For specific features, 89.9% of the variance in the number of SCRs and 72.7% of the variance in average SCR amplitudes were explained, with moderate to strong correlations (r=0.69 and r=0.61, respectively).

Table 5. Correspondence between MindWare recordings and wearable devices.

Device and Metric	β0j	U0j	β1j	U1j	SE	*P* value	Marginal R^2 (%)	Conditional R^2 (%)	Pearson *r*
Nuanic									
EDA[a]	1.25	13.73	1.17	1.10	0.15	<.001	46.6	99.3	0.78[b]
nSCR[c]	2.13	5.68	0.73	0.15	0.07	<.001	66.2	89.9	0.69
Amp_SCR[d]	0.29	0.08	0.90	0.57	0.14	<.001	33.9	72.7	0.61
Nowatch									
HR[e]	–0.16	20.82	1.01	0.007	0.01	<.001	95.5	96.1	0.87
MeanNN[f]	20.99	2368.16	0.96	0.003	0.02	<.001	94.1	94.8	0.85
SDNN[g]	24.70	372.79	0.54	0.11	0.05	<.001	42.3	77.9	0.55
RMSSD[h]	22.33	142.82	0.29	0.08	0.05	<.001	15.9	80.5	0.42
Oura Ring									
HR	1.87	132.73	0.98	0.02	0.02	<.001	87.4	91.2	0.68[b]
RMSSD	8.73	81.35	0.74	0.10	0.05	<.001	53.0	75.1	0.36[b]
Empatica									
HR	13.57	293.44	0.83	0.05	0.03	<.001	87.3	96.6	0.85[b]
RMSSD	28.43	370.59	0.22	0.07	0.04	<.001	17.8	90.3	0.41[b]

[a]EDA: electrodermal activity.
[b]Results from cross-correlation analysis.
[c]nSCR: number of skin conductance responses.
[d]Amp_SCR: average amplitude of skin conductance responses
[e]HR: heart rate.
[f]MeanNN: mean duration of normal-to-normal interbeat intervals.
[g]SDNN: SD of normal-to-normal intervals.
[h]RMSSD: root mean square of successive differences.

HR and MeanNN from Nowatch also demonstrated a strong association with lab ECG–derived metrics, with 96.1% and 94.8% of the variance explained, respectively, and strong correlations (r=0.87 and r=0.85, respectively; Figures 5 and 6). Associations for other HRV indices were weaker. SDNN and RMSSD showed acceptable model association but only weak to moderate correlations (r=0.55 and r=0.42, respectively). Oura rings showed a strong association with 25-second ECG window–derived HR (marginal R^2=87.4%, conditional R^2=91.2%) and good correlation (r=0.68) (Figure 7). HRV measures also showed good association (marginal R^2=53%, conditional R^2=75.1%), but their correlation with lab measures was weaker (r=0.36) (Figure 8).

Figure 5. Distribution of within-subject correlations between MindWare and Nowatch heart rate (HR). The top panel shows a histogram of correlation coefficients between MindWare and Nowatch HR measurements across all participants. The bottom panel shows overlaid HR traces from the participant whose correlation corresponds to the sample median (r=0.96). bpm: beats per minute.

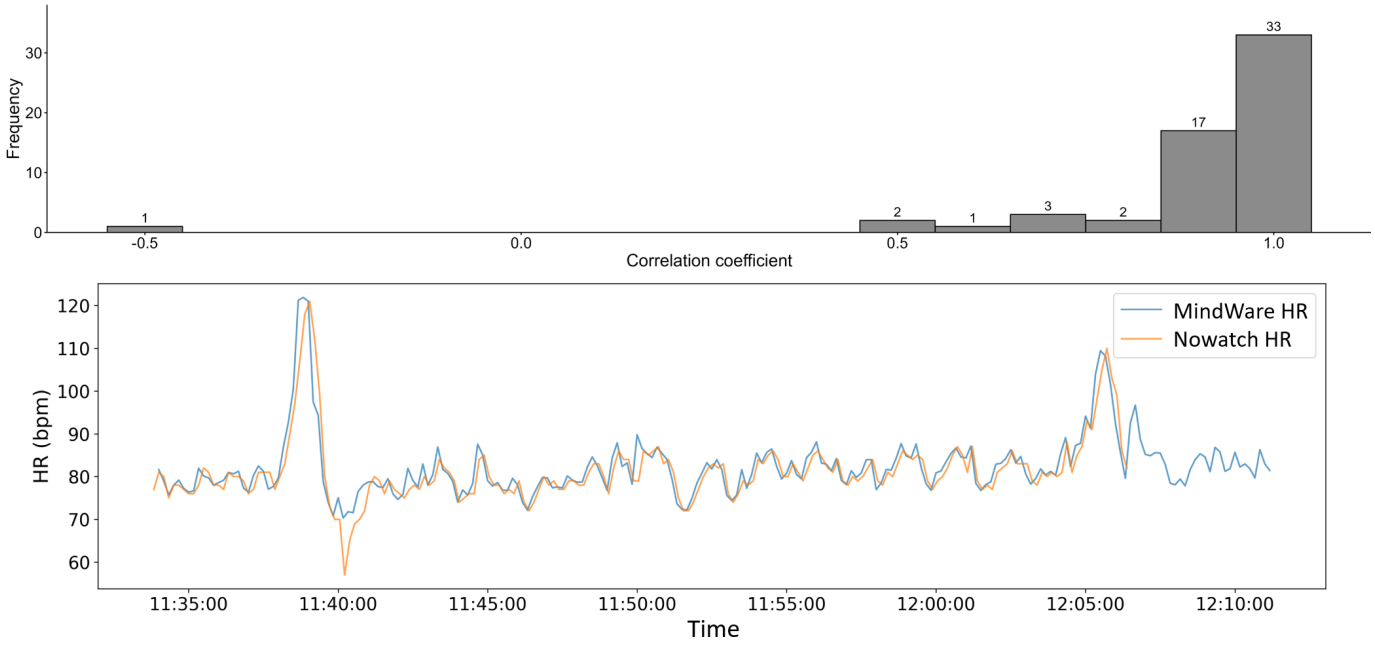


Figure 6. Distribution of within-subject correlations between MindWare and Nowatch interbeat interval (IBI). The top panel shows a histogram of correlation coefficients between MindWare and Nowatch IBI measurements across all participants. The bottom panel shows overlaid IBI traces from the participant whose correlation corresponds to the sample median (r=0.97). Only the first 200 samples are displayed to improve visual clarity.

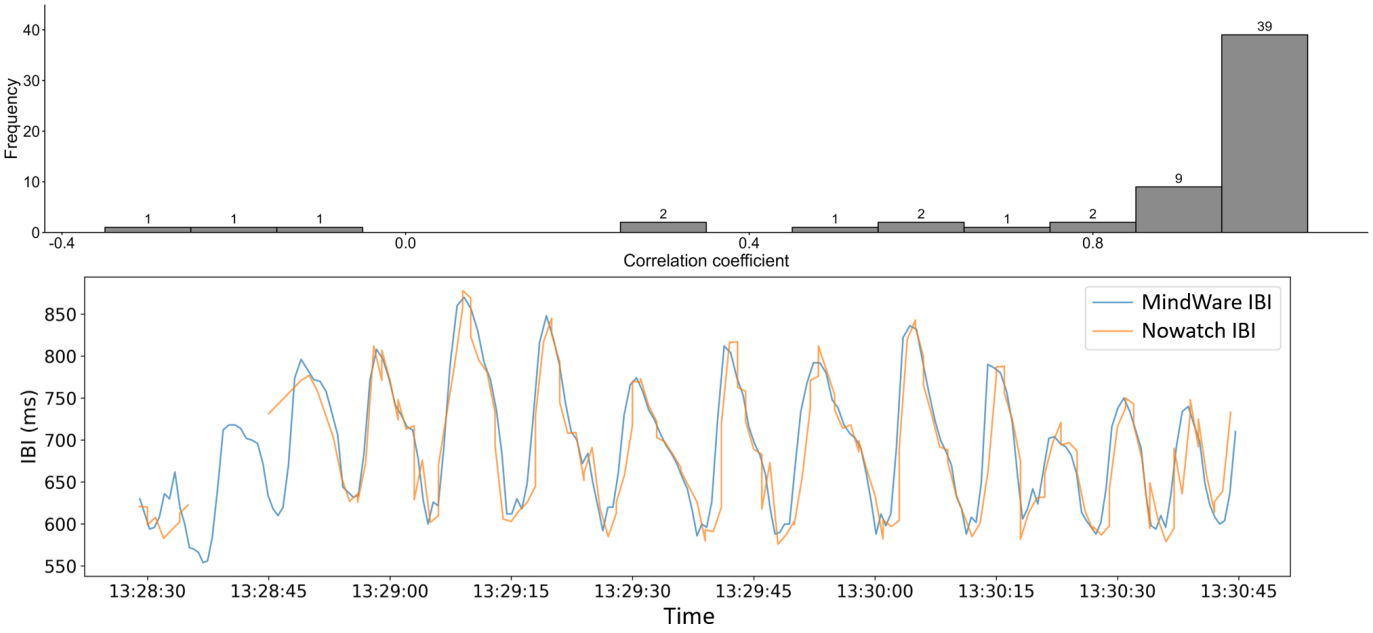


Figure 7. Distribution of within-subject correlations between MindWare and Oura heart rate (HR). The top panel shows a histogram of correlation coefficients between MindWare and Oura Ring "unguided session" HR measurements across all participants. The bottom panel shows overlaid HR traces from the participant whose correlation corresponds to the sample median (r=0.73). bpm: beats per minute.

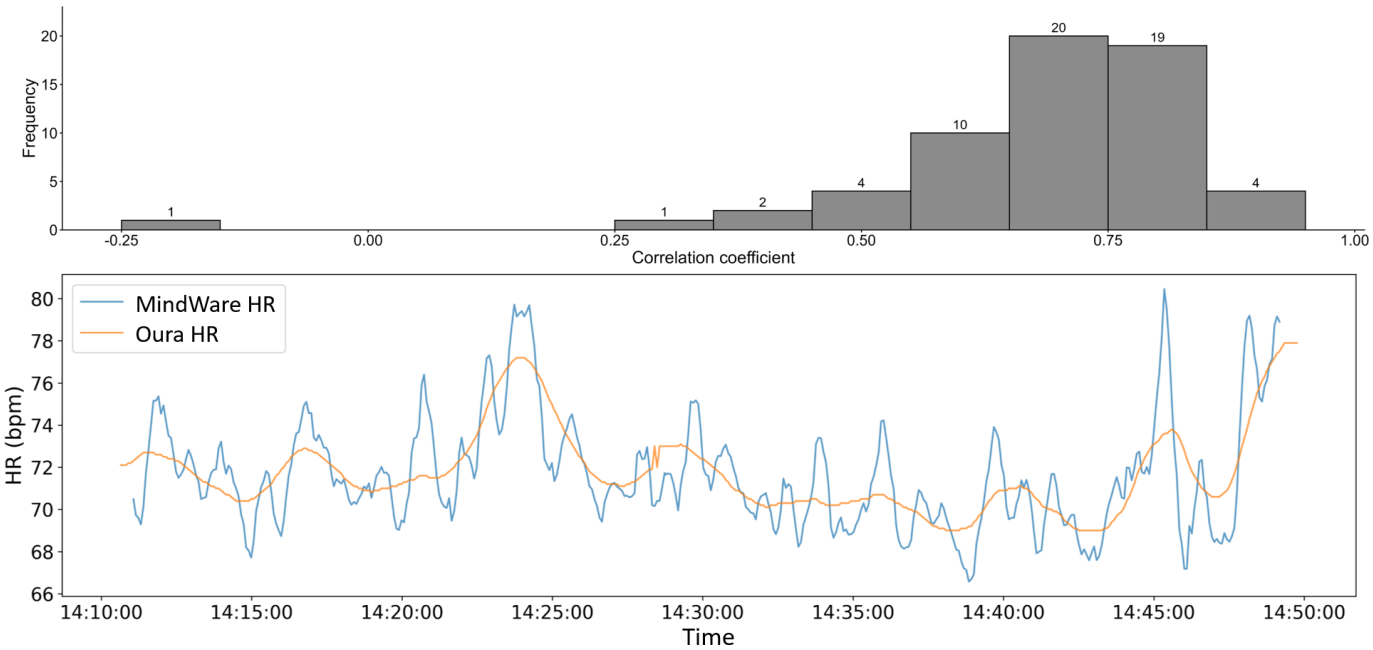
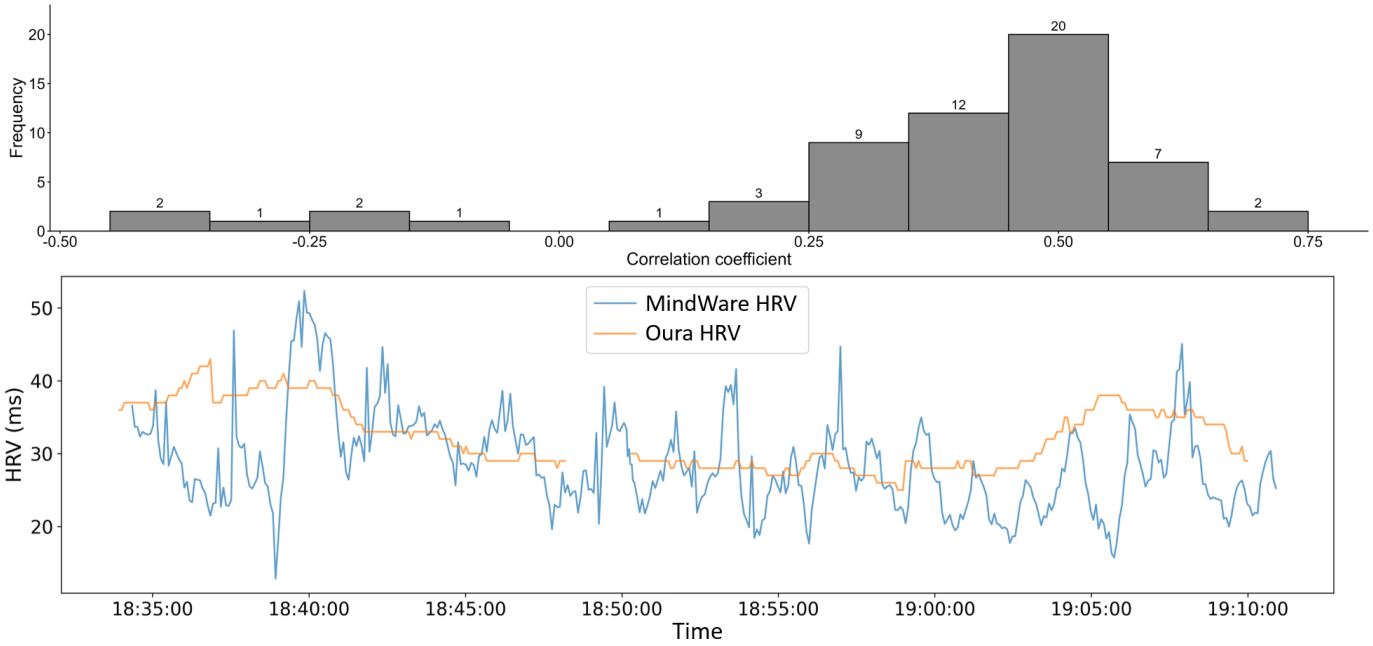


Figure 8. Distribution of within-subject correlations between MindWare and Oura heart rate variability (HRV). The top panel shows a histogram of correlation coefficients between MindWare and Oura Ring "unguided session" RMSSD (root mean square of successive differences) measurements across all participants. The bottom panel shows overlaid HRV traces from the participant whose correlation corresponds to the sample median (r=0.45).



Finally, HR from the Empatica EmbracePlus watches showed a strong association (marginal R^2=87.3%, conditional R^2=96.6%) and a strong correlation (r=0.85) (Figure 9). Although HRV also had a strong model fit (conditional R^2=90.3%), the substantially lower marginal R^2 (17.8%) (Figure 10) indicated that much of the explained variance was attributable to participant-level random effects rather than fixed effects alone. Consistent with this interpretation, cross-correlation indicated weaker synchrony (r=0.41) between Empatica RMSSD and lab ECG–derived RMSSD.

Figure 9. Distribution of within-subject correlations between MindWare and Empatica heart rate (HR). The top panel shows a histogram of correlation coefficients between MindWare and Empatica HR measurements across all participants. The bottom panel shows overlaid HR traces from the participant whose correlation corresponds to the sample median (r=0.95). bpm: beats per minute.

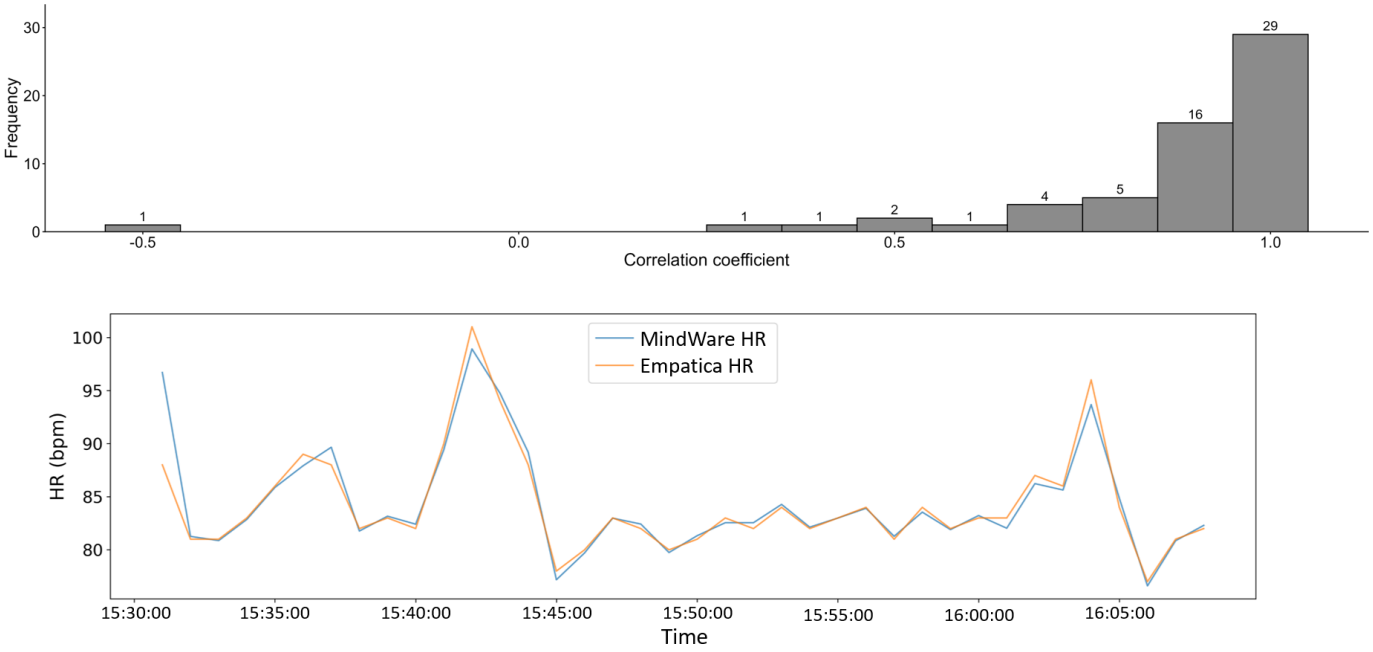
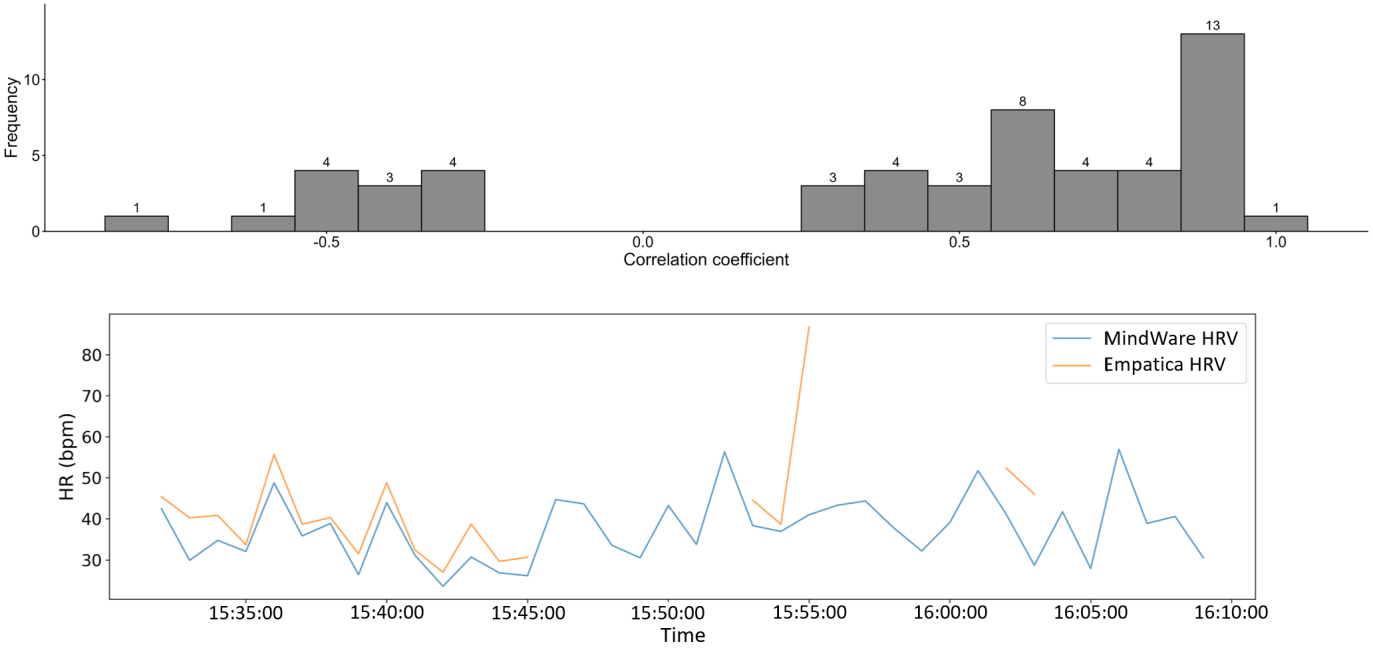


Figure 10. Distribution of within-subject correlations between MindWare and Empatica heart rate variability (HRV). The top panel shows a histogram of correlation coefficients between MindWare and Empatica HRV (root mean square of successive differences) measurements across all participants. The bottom panel shows overlaid HRV traces from the participant whose correlation corresponds to the sample median (r=0.85). bpm: beats per minute.



In addition, Bland-Altman analyses were performed to evaluate the agreement between the wearable devices and MindWare. The Bland-Altman plots are presented in Figure S3 (Multimedia Appendix 1), and the mean bias, 95% limits of agreement (LoA), MAE, and ICC results are summarized in Table 6. Overall, wearable devices demonstrated relatively small mean biases, indicating minimal systematic overestimation or underestimation of the reference measurements, except that Nowatch overestimated MeanNN and RMSSD and Empatica overestimated RMSSD. Average SCR amplitude in Nuanic and HR measured by Nowatch, Oura Ring, and Empatica exhibited relatively narrow LoA and small MAE, indicating good agreement with laboratory recordings. EDA and the number of SCRs from Nuanic exhibited relatively wide LoA despite low MAE, suggesting that most measurements were close to the reference values but some observations had larger differences that resulted in greater variability in agreement. Wider LoA and larger MAE were also observed for HRV metrics compared to other metrics. ICC results showed excellent agreement between wearable-derived and lab measurements for Nowatch HR and MeanNN, Oura Ring's HR, and Empatica's HR. Oura Ring's RMSSD showed good agreement, whereas Nowatch's SDNN and RMSSD showed moderate agreement. Other metrics, including all EDA metrics, had poor absolute agreement with the lab recordings.

Table 6. Bland-Altman analysis, mean absolute error (MAE), and intraclass correlation coefficient (ICC) comparing MindWare recordings and wearable devices.

Device and metric	Bias	Lower LoA[a]	Upper LoA	MAE	ICC (95% CI)
Nuanic					
EDA[b]	1.39	−16.96	19.74	4.00	0.34 (0.33-0.35)
nSCR[c]	0.53	−14.58	15.65	3.03	0.44 (0.37-0.50)
Amp_SCR[d]	−0.21	−1.54	1.12	0.40	0.32 (0.22-0.40)
Nowatch					
HR[e]	−0.66	−7.52	6.19	1.87	0.96 (0.95-0.97)
MeanNN[f]	8.74	−52.14	69.62	15.05	0.97 (0.96-0.97)
SDNN[g]	1.64	−36.79	40.06	10.67	0.73 (0.68-0.76)
RMSSD[h]	8.64	−29.57	46.85	13.17	0.59 (0.44-0.69)
Oura Ring					
HR	−0.15	−7.83	7.54	2.48	0.94 (0.94-0.94)
RMSSD	−0.12	−21.64	21.41	7.17	0.87 (0.86-0.87)
Empatica					
HR	−0.75	−7.65	6.16	1.42	0.96 (0.95-0.96)
RMSSD	19.38	−49.19	87.95	19.77	0.47 (0.25-0.62)

Device and metric	Bias	Lower LoA[a]	Upper LoA	MAE	ICC (95% CI)

[a]LoA: limits of agreement.
[b]EDA: electrodermal activity.
[c]nSCR: number of skin conductance responses.
[d]Amp_SCR: average amplitude of skin conductance responses.
[e]HR: heart rate.
[f]MeanNN: mean duration of normal-to-normal interbeat intervals.
[g]SDNN: SD of normal-to-normal intervals.
[h]RMSSD: root mean square of successive differences.

Across devices, Nuanic demonstrated a strong association with laboratory EDA and reliably captured the direction of task-related changes, although the number and amplitudes of SCR showed lower correspondence and sensitivity. Nowatch HR and MeanNN closely matched the lab ECG recordings and consistently tracked task-induced increases in HR, while other HRV indices (ie, SDNN and RMSSD) showed weaker correspondence and did not consistently reflect changes across task phases. Oura Ring and Empatica HR also aligned well with ECG-derived HR, but HRV showed a lower association.

Discussion

The present study evaluated the validity of 4 wearable sensor devices (Nowatch, Empatica, Nuanic, and Oura Ring) against laboratory-based gold standard measurements in HR, HRV, and EDA metrics during rest periods and stress induction tasks under controlled laboratory conditions with minimized movement.

Nuanic demonstrated strong performance at both the signal and event levels, where signal-level performance reflects correspondence with laboratory-based EDA recordings and event-level performance reflects sensitivity to experimentally induced stress-related changes. It was surprising to observe a good correlation at the signal level because measurements from wearable devices tend to vary significantly because of sensor type and measurement locations [8]. Bland-Altman analysis further supported the overall agreement between Nuanic and laboratory EDA for SCR amplitude. For the number of SCRs, although most Nuanic measurements were close to the lab reference values, a subset of observations showed larger discrepancies, contributing to greater variability. Consistent with this finding, there was higher variability (reflected as the SD) in Nuanic signals compared to the gold standard lab recordings. After the data were decomposed into tonic and phasic components to extract the nSCR and Amp_SCR data, both variation and correspondence decreased. We also noticed that more SCRs were detected during the resting phases by Nuanic compared to lab EDA, which could be due to increased sensitivity of the ring owing to different sensor types and measurement locations or because Ledalab decomposition applied to low-frequency data (16 Hz) resulted in a noisier phasic driver.

HR and MeanNN from Nowatch also exhibited strong validity and task sensitivity, while other HRV indices showed a weaker association with the lab reference signals. At the signal level, similar to Nuanic, both the correlation and LMM results indicated that the HR and MeanNN overlay well with the HR and MeanNN extracted from the ECG recording; however, larger discrepancies were observed in SDNN and RMSSD. Bland-Altman analysis showed relatively small mean biases for HR and MeanNN, whereas SDNN and RMSSD showed greater variability and measurement error. Both SDNN and RMSSD are much more sensitive to artifacts and missing data compared to MeanNN, and movement artifacts are heightened at ambulatory settings [28,29]. Additionally, past studies have shown that SDNN is less reliable in recording windows shorter than 2 minutes [30], which is relevant because our stress induction tasks were shorter than this duration. Because of these limitations, the results for SDNN on responsiveness to task-induced stress should be interpreted with caution, whereas the results for MeanNN and RMSSD should be considered relatively robust.

Previous validation studies on the Oura Ring have investigated HR and HRV metrics during sleep or during short sessions when people were sitting still [31,32]. The present study used the “unguided session” function and successfully extracted the HR and HRV metrics during the daytime; however, the method is not realistic to implement when measuring in a free-living setting due to the amount of restriction in movement needed. The extracted results showed good correspondence for HR but a relatively weak correlation for HRV (RMSSD). The algorithm used by Oura to calculate HR and RMSSD was not available; however, data observation suggests that Oura Ring’s outputs are generated by taking the average of segments of unknown duration (see Figures 7 and 8 for examples). Since we could not replicate the data processing procedure from Oura, the current data showed a weaker correspondence between devices and worse validation results. Sensor location may also cause discrepancies in the correspondence between wearable measurements and the laboratory reference signals; however, past studies have shown that a finger is a more optimal location than the wrist to record the pulse rate due to its rich vascularization and optimal skin thickness [33,34]. Therefore, we should expect to see better results from the Oura Ring than from wrist-worn watches. According to the API documentation from Oura, the Oura Ring provides higher-resolution HR and HRV readings during sleep when movement is limited [35] and could potentially generate better results when data segments are averaged over longer intervals (eg, every 5 minutes during sleep).

Extensive validation research has been done on Empatica's previous model, E4 [8,36,37], but very few validation studies have been conducted on the newer model EmbracePlus. Similarly, Nowatch is relatively new to the market, and to the best of our knowledge, no studies to date have formally validated the watch's measurement against laboratory-grade reference measures. The Empatica watches and Nowatch have both EDA and PPG sensors, which is ideal when recording multiple-model stress indices. However, our current methodology was not able to record reliable EDA readings in Empatica, possibly due to the short recording duration. However, we explicitly built in a warm-up period, where our research assistant wore both Nowatch and the Empatica watch for 15 minutes before participants came in, as recommended by both manufacturers, to ensure that the sensors were "warmed up" before the watch was transferred to a participant's wrist. Another possible explanation is the lower sensitivity of the wrist EDA sensor to physiological stress compared to palmar EDA sensor due to the lower density of eccrine sweat glands on the wrist [14]. This reduced sensitivity may limit the ability of wrist EDA sensors to detect subtle changes in psychological manipulations. On the other hand, Nowatch's EDA could not be recorded due to the recording setting. This was a software limitation that we had noticed and requested the manufacturer to address. Further research needs to be cautious in accurately collecting EDA in different settings, both in the lab and in real life. We analyzed the processed HR and HRV outputs from Empatica's proprietary algorithms and found strong validity in HR readings. HRV readings were largely compromised by motion artifacts, and about half of the recorded data were not valid, even though the study was done in a relatively stationary condition. This suggests that while the device performs well for basic cardiovascular measures such as HR, more complex metrics such as RMSSD may be more sensitive to noise and preprocessing assumptions embedded in proprietary pipelines. Given that HRV is widely used as an index of autonomic reactivity and stress regulation, these limitations found across devices have important implications for interpretation, particularly in studies relying on ambulatory protocols. Importantly, Empatica does provide access to raw PPG data, which allows for an independent computation of IBI and HRV metrics following appropriate artifact correction and signal processing. This may offer a viable pathway for improving measurement in future studies and mitigating limitations associated with proprietary HRV outputs.

The present study had limitations. First, the time stamps were not perfectly matched with the stimulus onset and offset time stamps due to different sampling rates. Time shifts were also present in Nuanic during recording. This is a device limitation that the manufacturer acknowledged and is actively working to resolve; however, it is currently a barrier to data analysis if fine time resolution is required and in real-world settings where no standard reference time is available. In our study, significant time misalignments were observed in Nuanic EDA data and corrected by visual inspection. Although this approach reduces the influence of the technical issue, it may also introduce potential observer bias and should be considered when interpreting the results. We used cross-correlation analysis on time series data (EDA and Oura-derived HR and HRV metrics) to minimize the impact of the temporal lags between signals. Second, our study validated the sensors' performance in a laboratory setting where motion was limited compared to an ambulatory setting. Physiological recordings such as PPG and EDA are prone to motion artifacts [38]. Even though Nowatch and the Oura Ring contain proprietary algorithms to detect and suppress such artifacts, the validity of the data in motion has not been proved. Future studies should evaluate the devices' performance and data validity in motion. Third, the majority of participants were female, consistent with the sex distribution of psychology majors at our institution and others. Further research is needed to assess for possible sex differences in physiological responses to stress, including EDA dynamics and HRV, as well as in the validity and reliability of wearable-derived measures. The study was conducted in healthy young adults (university students). Neurological disorders, cardiac disorders, and medications known to affect the autonomic nervous system or cardiovascular system were the exclusion criteria. Therefore, the results of the present study should not be assumed to generalize to clinical populations, particularly those with underlying medical conditions or taking medications that may affect physiological responses. Previous research has also shown that optical sensors such as PPG sensors might be affected by skin tones [39,40]. Because the majority of our participants had lighter skin tones (85.48% East Asian, Southeast Asian, or White), further validation in more diverse populations is needed to assess the generalizability of these findings across different skin tones.

Another limitation is that our experimental stress induction tasks produced differing levels of subjective stress and physiological response (across the devices), but we cannot be certain why. Images from the OASIS selected to be highly arousing negative and positive stimuli were not successful in eliciting physiological responses including EDA and HR-related metrics. We observed the strongest stress-related responses during PASAT and SSST. Both of these tasks required speaking aloud, which could have contributed to EDA and HRV changes. In other words, the stronger physiological and behavioral responses observed during PASAT and SSST may reflect the combined effects of cognitive demand, verbal output, and perceived social pressure.

Finally, Ledalab's continuous decomposition analysis was used to decompose EDA into number of SCRs and SCR amplitude, which relies on a mathematical model of the sudomotor response, and the resulting estimates may be influenced by the assumptions and parameters of the model [23]. As mentioned above, it is also sensitive to the temporal resolution of the input EDA signal, and differences in the sampling frequency affect the detection of individual responses, including the number of SCRs and SCR amplitude. Therefore, differences observed in EDA-derived metrics may reflect not only the differences between the devices but also the differences in the results produced by Ledalab when processing signals with different temporal resolutions.

Taken together, the findings indicate that the wearable devices evaluated were able to capture some aspects of physiological stress reactivity under controlled laboratory conditions, with performance varying by device and metric. While certain measures (Nuanic's EDA and Nowatch's HR and MeanNN) showed acceptable agreement with laboratory-based systems, other metrics underperformed and should be used with caution in a noncontrolled environment.

The present study provides valuable preliminary evidence supporting the physiological measurements from the Nuanic ring, Nowatch, and Oura Ring in controlled laboratory settings while participants were stationary. These findings contribute to the growing literature on the validity of consumer-grade wearables and highlight both their potential and current constraints for multimodal research and clinical applications. Multimodal monitoring of physiological measures has potential for assessing stress and cognitive states in ecological and operational settings, including safety-critical environments [41-43]. However, previous research comparing the performance of wearable devices across settings demonstrated reduced accuracy in ambulatory settings [44,45]. Continued validation work—particularly in real-world environments—will be essential to ensure the reliability of these devices for capturing physiological markers of stress and arousal beyond the laboratory.

Acknowledgments

We thank Yaoyue Feng for helping assess the reliability of Nuanic's time shifts.

Generative AI Use

Grammarly was used as a grammar checker to improve the clarity of language in the first manuscript draft. The suggested changes were reviewed and verified by the authors, who take full responsibility for the final manuscript.

Funding

The study was supported by the VGH & UBC Hospital Foundation and the Canada Foundation for Innovation (CFI) John R Evans Leaders Fund. The PhD student JG is supported by a Mitacs Accelerate Fellowship (IT37387).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: NDS, JG, CLP, RMT, ATH
Data curation: JG
Formal analysis: JG
Methodology: NDS, JG, CLP, RMT, ATH
Supervision: NDS
Visualization: JG
Writing – original draft: JG
Writing – review & editing: NDS, CLP, RMT, ATH

Conflicts of Interest

None declared.

Multimedia Appendix 1

Schematic depicting the Open Affective Standardized Image Set image viewing task and tables depicting heart rate comparison using different electrocardiogram window sizes, linear mixed model and fitted results between devices, Bland-Altman plots comparing wearables and the reference device, and repeated measures ANOVA results.
[DOCX File (Microsoft Word File), 4456 KB-Multimedia Appendix 1]

References

1. Knight S, Lipoth J, Namvari M, et al. The accuracy of wearable photoplethysmography sensors for telehealth monitoring: a scoping review. Telemed J E Health. Jun 2023;29(6):813-828. [doi: 10.1089/tmj.2022.0182] [Medline: 36288566]
2. Charlton PH, Marozas V. Wearable photoplethysmography devices. In: Allen J, Kyriacou PA, editors. Photoplethysmography: Technology, Signal Analysis and Applications. Academic Press (Elsevier); 2021:401-439. [doi: 10.1016/B978-0-12-823374-0.00011-6]
3. Posada-Quintero HF, Chon KH. Innovations in electrodermal activity data collection and signal processing: a systematic review. Sensors (Basel). Jan 15, 2020;20(2):479. [doi: 10.3390/s20020479] [Medline: 31952141]
4. Ghiasi S, Greco A, Barbieri R, Scilingo EP, Valenza G. Assessing autonomic function from electrodermal activity and heart rate variability during cold-pressor test and emotional challenge. Sci Rep. Mar 25, 2020;10(1):5406. [doi: 10.1038/s41598-020-62225-2] [Medline: 32214158]

5. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig. Mar 2018;15(3):235-245. [doi: 10.30773/pi.2017.08.17] [Medline: 29486547]
6. Peake JM, Kerr G, Sullivan JP. A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. Front Physiol. 2018;9:743. [doi: 10.3389/fphys.2018.00743] [Medline: 30002629]
7. Miller DJ, Sargent C, Roach GD. A validation of six wearable devices for estimating sleep, heart rate and heart rate variability in healthy adults. Sensors (Basel). Aug 22, 2022;22(16):6317. [doi: 10.3390/s22166317] [Medline: 36016077]
8. Milstein N, Gordon I. Validating measures of electrodermal activity and heart rate variability derived from the Empatica E4 utilized in research settings that involve interactive dyadic states. Front Behav Neurosci. 2020;14:148. [doi: 10.3389/fnbeh.2020.00148] [Medline: 33013337]
9. Costantini S, Chiappini M, Malerba G, et al. Wrist-worn sensor validation for heart rate variability and electrodermal activity detection in a stressful driving environment. Sensors (Basel). Oct 12, 2023;23(20):8423. [doi: 10.3390/s23208423] [Medline: 37896517]
10. Pakarinen T, Pietila J, Nieminen H. Prediction of self-perceived stress and arousal based on electrodermal activity. In: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2019:2191-2195. [doi: 10.1109/EMBC.2019.8857621]
11. Torniainen J, Cowley B, Henelius A, Lukander K, Pakarinen S. Feasibility of an electrodermal activity ring prototype as a research tool. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2015:6433-6436. [doi: 10.1109/EMBC.2015.7319865]
12. van Lier HG, Pieterse ME, Garde A, et al. A standardized validity assessment protocol for physiological signals from wearable technology: methodological underpinnings and an application to the E4 biosensor. Behav Res Methods. Apr 2020;52(2):607-629. [doi: 10.3758/s13428-019-01263-9] [Medline: 31290128]
13. Sartor F, Papini G, Cox LGE, Cleland J. Methodological shortcomings of wrist-worn heart rate monitors validations. J Med Internet Res. Jul 2, 2018;20(7):e10108. [doi: 10.2196/10108] [Medline: 29967000]
14. van der Mee DJ, Gevonden MJ, Westerink J, de Geus EJC. Validity of electrodermal activity-based measures of sympathetic nervous system activity from a wrist-worn device. Int J Psychophysiol. Oct 2021;168:52-64. [doi: 10.1016/j.ijpsycho.2021.08.003] [Medline: 34418464]
15. Tombaugh TN. A comprehensive review of the Paced Auditory Serial Addition Test (PASAT). Arch Clin Neuropsychol. Jan 2006;21(1):53-76. [doi: 10.1016/j.acn.2005.07.006] [Medline: 16290063]
16. Cellini N, Grondin S, Stablum F, Sarlo M, Mioni G. Psychophysiological stress influences temporal accuracy. Exp Brain Res. Sep 2023;241(9):2229-2240. [doi: 10.1007/s00221-023-06676-9] [Medline: 37530787]
17. Brouwer AM, Hogervorst MA. A new paradigm to induce mental stress: the Sing-a-Song Stress Test (SSST). Front Neurosci. 2014;8(224):224. [doi: 10.3389/fnins.2014.00224] [Medline: 25120425]
18. Carpenter J, Sherman MT, Kievit RA, Seth AK, Lau H, Fleming SM. Domain-general enhancements of metacognitive ability through adaptive training. J Exp Psychol Gen. Jan 2019;148(1):51-64. [doi: 10.1037/xge0000505] [Medline: 30596440]
19. Kurdi B, Lozano S, Banaji MR. Introducing the Open Affective Standardized Image Set (OASIS). Behav Res. Apr 2017;49(2):457-470. [doi: 10.3758/s13428-016-0715-3] [Medline: 26907748]
20. Electrode Placement: ECG, Impedance Cardiography, GSR. MindWare Technologies Ltd; 2016. URL: https://support.mindwaretech.com/wp-content/uploads/2016/04/Electrode_Placement_11x85.pdf [Accessed 2026-11-09]
21. Peirce JW. Generating stimuli for neuroscience using PsychoPy. Front Neuroinform. 2008;2:10. [doi: 10.3389/neuro.11.010.2008] [Medline: 19198666]
22. Vanderwal T, Kelly C, Eilbott J, Mayes LC, Castellanos FX. Inscapes: a movie paradigm to improve compliance in functional magnetic resonance imaging. Neuroimage. Nov 15, 2015;122:222-232. [doi: 10.1016/j.neuroimage.2015.07.069] [Medline: 26241683]
23. Benedek M, Kaernbach C. A continuous measure of phasic electrodermal activity. J Neurosci Methods. Jun 30, 2010;190(1):80-91. [doi: 10.1016/j.jneumeth.2010.04.028] [Medline: 20451556]
24. Kleckner IR, Jones RM, Wilder-Smith O, et al. Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data. IEEE Trans Biomed Eng. Jul 2018;65(7):1460-1467. [doi: 10.1109/TBME.2017.2758643] [Medline: 28976309]
25. van Dooren M, de Vries J, Janssen JH. Emotional sweating across the body: comparing 16 different skin conductance measurement locations. Physiol Behav. May 15, 2012;106(2):298-304. [doi: 10.1016/j.physbeh.2012.01.020] [Medline: 22330325]
26. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav Res Methods. Aug 2021;53(4):1689-1696. [doi: 10.3758/s13428-020-01516-y] [Medline: 33528817]

27. Unguided sessions. Oura Ring. 2025. URL: https://support.ouraring.com/hc/en-us/articles/360025584753-Unguided-Sessions#h_01FW7E6TY9DRWTJEDSFN428BY0 [Accessed 2026-10-31]
28. Bourdillon N, Yazdani S, Vesin JM, Schmitt L, Millet GP. RMSSD is more sensitive to artifacts than frequency-domain parameters: implication in athletes' monitoring. J Sports Sci Med. Jun 2022;21(2):260-266. [doi: 10.52082/jssm.2022.260] [Medline: 35719238]
29. Benchekroun M, Chevallier B, Zalc V, Istrate D, Lenne D, Vera N. The impact of missing data on heart rate variability features: a comparative study of interpolation methods for ambulatory health monitoring. IRBM. Aug 2023;44(4):100776. [doi: 10.1016/j.irbm.2023.100776]
30. Munoz ML, van Roon A, Riese H, et al. Validity of (ultra-)short recordings for heart rate variability measurements. PLoS ONE. 2015;10(9):e0138921. [doi: 10.1371/journal.pone.0138921] [Medline: 26414314]
31. Cao R, Azimi I, Sarhaddi F, et al. Accuracy assessment of Oura Ring nocturnal heart rate and heart rate variability in comparison with electrocardiography in time and frequency domains: comprehensive analysis. J Med Internet Res. Jan 18, 2022;24(1):e27487. [doi: 10.2196/27487] [Medline: 35040799]
32. 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. May 7, 2020;41(4):04NT01. [doi: 10.1088/1361-6579/ab840a] [Medline: 32217820]
33. Nardelli M, Vanello N, Galperti G, Greco A, Scilingo EP. Assessing the quality of heart rate variability estimated from wrist and finger PPG: a novel approach based on cross-mapping method. Sensors (Basel). Jun 2, 2020;20(11):3156. [doi: 10.3390/s20113156] [Medline: 32498403]
34. Longmore SK, Lui GY, Naik G, Breen PP, Jalaludin B, Gargiulo GD. A comparison of reflective photoplethysmography for detection of heart rate, blood oxygen saturation, and respiration rate at various anatomical locations. Sensors (Basel). Apr 19, 2019;19(8):1874. [doi: 10.3390/s19081874] [Medline: 31010184]
35. Oura API v2 documentation. Ōura Health Oy. 2026. URL: https://cloud.ouraring.com/v2/docs [Accessed 2026-10-31]
36. Stuyck H, Dalla Costa L, Cleeremans A, Van den Bussche E. Validity of the Empatica E4 wristband to estimate resting-state heart rate variability in a lab-based context. Int J Psychophysiol. Dec 2022;182:105-118. [doi: 10.1016/j.ijpsycho.2022.10.003] [Medline: 36252721]
37. Schuurmans AAT, de Looff P, Nijhof KS, et al. Validity of the Empatica E4 wristband to measure heart rate variability (HRV) parameters: a comparison to electrocardiography (ECG). J Med Syst. Sep 23, 2020;44(11):190. [doi: 10.1007/s10916-020-01648-w] [Medline: 32965570]
38. Pollreisz D, TaheriNejad N. Detection and removal of motion artifacts in PPG signals. Mobile Netw Appl. Apr 2022;27(2):728-738. [doi: 10.1007/s11036-019-01323-6]
39. Hung SH, Serwa K, Rosenthal G, Eng JJ. Validity of heart rate measurements in wrist-based monitors across skin tones during exercise. PLoS One. 2025;20(2):e0318724. [doi: 10.1371/journal.pone.0318724] [Medline: 39928630]
40. Koerber D, Khan S, Shamsheri T, Kirubarajan A, Mehta S. Accuracy of heart rate measurement with wrist-worn wearable devices in various skin tones: a systematic review. J Racial Ethn Health Disparities. Dec 2023;10(6):2676-2684. [doi: 10.1007/s40615-022-01446-9] [Medline: 36376641]
41. Ronca V, Martinez-Levy AC, Vozzi A, et al. Wearable technologies for electrodermal and cardiac activity measurements: a comparison between Fitbit Sense, Empatica E4 and Shimmer GSR3. Sensors (Basel). Jun 23, 2023;23(13):5847. [doi: 10.3390/s23135847] [Medline: 37447697]
42. Borghini G, Bandini A, Orlandi S, et al. Stress assessment by combining neurophysiological signals and radio communications of air traffic controllers. Annu Int Conf IEEE Eng Med Biol Soc. Jul 2020;2020:851-854. [doi: 10.1109/EMBC44109.2020.9175958] [Medline: 33018118]
43. Ronca V, Castagneto Gissey L, Bellini MI, et al. Multi-method characterization of neurophysiological and biological stress responses in surgical teams during real surgical procedures. Front Neuroergon. 2026;7:1702748. [doi: 10.3389/fnrgo.2026.1702748] [Medline: 41788916]
44. Hu X, Sgherza TR, Nothrup JB, Fresco DM, Naragon-Gainey K, Bylsma LM. From lab to life: evaluating the reliability and validity of psychophysiological data from wearable devices in laboratory and ambulatory settings. Behav Res Methods. Oct 2024;56(7):1-20. [doi: 10.3758/s13428-024-02387-3] [Medline: 38528248]
45. Sinichi M, Gevonden MJ, Krabbendam L. Quality in question: assessing the accuracy of four heart rate wearables and the implications for psychophysiological research. Psychophysiology. Feb 2025;62(2):e70004. [doi: 10.1111/psyp.70004] [Medline: 39905563]

Abbreviations

Amp_SCR: average amplitude of skin conductance responses
ECG: electrocardiogram
EDA: electrodermal activity

HR: heart rate
HRV: heart rate variability
IBI: interbeat interval
ICC: intraclass correlation coefficient
LMM: linear mixed effects model
LoA: 95% limits of agreement
MAE: mean absolute error
MeanNN: mean duration of normal-to-normal interbeat intervals
nSCR: number of skin conductance responses
OASIS: Open Affective Standardized Image Set
PASAT: Paced Auditory Serial Addition Test
PPG: photoplethysmography
RMSSD: root mean square of successive differences
SCR: skin conductance response
SDNN: SD of normal-to-normal intervals
SSST: Sing-a-Song Stress Test

Edited by Javad Sarvestan; peer-reviewed by Hafiz Muhammad Salman Ajmal, Vincenzo Ronca; submitted 27.May.2026; initial comments to author 13.Jul.2026; final revised version received 16.Oct.2026; accepted 19.Oct.2026; published 20.Nov.2026

Please cite as:
Guo J, Pollock CL, Todd RM, Herdman AT, Silverberg ND
Validation of Heart Rate– and Electrodermal Activity–Related Signals From Commercial-Grade Wearable Smartwatches and Rings in Experimentally Induced Stress: Laboratory-Based Study
JMIR Biomed Eng 2026;11:e101911
URL: https://biomedeng.jmir.org/2026/1/e101911
doi: 10.2196/101911

© Jiaer Guo, Courtney L Pollock, Rebecca M Todd, Anthony T Herdman, Noah D Silverberg. Originally published in JMIR Biomedical Engineering (http://biomsedeng.jmir.org), 20.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Biomedical Engineering, is properly cited. The complete bibliographic information, a link to the original publication on https://biomedeng.jmir.org/, as well as this copyright and license information must be included.