# *Framework for selecting and benchmarking mobile devices in psychophysiological research*

**Ian R. Kleckner, Mallory J. Feldman, Matthew S. Goodwin & Karen S. Quigley**


**Behavior Research Methods**

e-ISSN 1554-3528

Behav Res
DOI 10.3758/s13428-020-01438-9

Springer

Behavior Research Methods
https://doi.org/10.3758/s13428-020-01438-9

# Framework for selecting and benchmarking mobile devices in psychophysiological research

Ian R. Kleckner[1] · Mallory J. Feldman[2,3] · Matthew S. Goodwin[2] · Karen S. Quigley[2,4]



## Abstract

Commercially available consumer electronics in (smartwatches and wearable biosensors) are increasingly enabling acquisition of peripheral physiological and physical activity data inside and outside of laboratory settings. However, there is scant literature available for selecting and assessing the suitability of these novel devices for scientific use. To overcome this limitation, the current paper offers a framework to aid researchers in choosing and evaluating wearable technologies for use in empirical research. Our seven-step framework includes: (1) identifying signals of interest; (2) characterizing intended use cases; (3) identifying study-specific pragmatic needs; (4) selecting devices for evaluation; (5) establishing an assessment procedure; (6) performing qualitative and quantitative analyses on resulting data; and, if desired, (7) conducting power analyses to determine sample size needed to more rigorously compare performance across devices. We illustrate the application of the framework by comparing electrodermal, cardiovascular, and accelerometry data from a variety of commercial wireless sensors (Affectiva Q, Empatica E3, Empatica E4, Actiwave Cardio, Shimmer) relative to a well-validated, wired MindWare laboratory system. Our evaluations are performed in two studies ($N = 10$, $N = 11$) involving psychometrically sound, standardized tasks that include physical activity and affect induction. After applying our framework to this data, we conclude that only some commercially available consumer devices for physiological measurement are capable of wirelessly measuring peripheral physiological and physical activity data of sufficient quality for scientific use cases. Thus, the framework appears to be beneficial at suggesting steps for conducting more systematic, transparent, and rigorous evaluations of mobile physiological devices prior to deployment in studies.

**Keywords** Monitoring · Ambulatory · Accelerometry · Benchmarking · Heart rate · Electrodermal activity · Psychophysiology · Affect · Stress

Recent advances in miniaturized hardware and wearable technology are enabling the use of smartwatches and mobile sensors to measure cardiovascular, electrodermal, and accelerometric data in empirical studies (Goodwin et al., 2008; Mukhopadhyay, 2015; Patel et al., 2012; Strangman et al., 2018). These telemetric devices (i.e., wearable equipment that often contain multiple sensors wherein different dependent variables can be measured) have several appealing affordances for studies that emphasize longitudinal or within-subject designs. Laboratory experiments typically explore inter-individual variability across a restricted number of scenarios, within tightly controlled environments, using relatively homogenous samples (Molenaar, 2004). Although useful for experimental control and inference, these paradigms necessarily restrict the study of context, temporal dynamics, and heterogeneity within and across individuals (Conner et al., 2009; Fisher et al., 2018; Patel et al., 2012).

In contrast, telemetric devices can capture intensive longitudinal data across time and real-world contexts both within and across individuals (Myrtek, 2004). Such capabilities are advantageous because they are more likely to capture events that are rare and unpredictable (e.g., panic attacks or cardiac

Shared first authorship Ian R. Kleckner and Mallory J. Feldman

Shared senior authorship Matthew S. Goodwin and Karen S. Quigley

✉ Ian R. Kleckner
Ian_Kleckner@URMC.Rochester.edu

[1] Cancer Control Unit, Department of Surgery, Department of Neuroscience, University of Rochester Medical Center, 265 Crittenden Blvd, Box CU 420658, Rochester, NY 14642, USA

[2] Northeastern University, Boston, MA, USA

[3] University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[4] Edith Nourse Rogers Memorial (VA) Medical Center, Bedford, MA, USA

Published online: 03 August 2020

arrhythmias; Leibold & Schruers, 2018; Mittal, Movsowitz, & Steinberg, 2011; Mittal et al., 2011), events that unfold over longer periods of time (e.g., sleep across days or metabolic changes with physical activity; Gao, Brooks, & Klonoff, 2018; Sano, Picard, & Stickgold, 2014), or salient events that may be unethical to elicit experimentally (e.g., receiving news about the death of a loved one; Wilhelm & Grossman, 2010). Ambulatory physiological recordings have also demonstrated utility performing dynamic assessments of symptoms over time in patients with cancer (Savard et al., 2013), Parkinson's disease (Moore et al., 2008), autism spectrum disorder (Goodwin et al., 2019), borderline personality disorder (Ebner-Priemer et al., 2008), and seizures (Michel et al., 2015). Finally, telemetric devices are also beginning to be used to deliver interventions to treat symptoms or disease (e.g., exercise interventions for patients with cancer; Schaffer et al., 2019).

Despite the potential, availability, and popularity of telemetric devices, development of mobile sensors consistently outpaces the rate of independent validation of these technologies against gold-standard, research-grade devices (Peake et al., 2018). The fact that validation efforts lag behind hardware development is a critical challenge given the importance that scientists, practitioners, and other conscientious users place on measurement fidelity. Moreover, traditional validation studies are often constrained by scope and context dependence. The acquisition of valid data from a telemetric device depends on a number of factors, including user experience and signal quality. Extant validation studies typically limit their assessment to one of these two categories, focusing exclusively on *either* user experience (e.g., Beaukenhorst et al. 2020) *or* signal quality. Additionally, those that focus on signal quality tend to emphasize *either* qualitative measures (e.g., McCarthy et al., 2016) *or* quantitative measures (e.g., Kasos et al., 2019; Straiton et al., 2018; van Lier et al., 2019; Weippert et al., 2010). While each of these categories of validation are informative and useful in their own right, variability in approaches can be intimidating for newcomers interested in utilizing ambulatory measurement in their research.

While science benefits from published guidelines, they too can be limited in scope by focusing on specific signals, statistical methods, or analytic decision criteria (Parati et al., 2010, 2014; van Lier et al., 2019). Rarely is one set of criteria sufficient for establishing validity and utility in science. In the present paper we attempt to address these obstacles by offering a multi-level, general-purpose framework for selecting, testing, comparing, and documenting the performance of wearable peripheral physiological devices for specific use cases. In so doing, we hope to deliver a more comprehensive conceptual scheme for establishing sufficient accuracy, precision, *and* feasibility of these emerging research tools in scientific studies.

Sufficient *accuracy* can be demonstrated when signals from a new sensor are shown to be comparable to those collected by a 'gold-standard' measurement of the same outcome variable. However, and critically, what is considered "sufficiently accurate" depends on both the type of data being collected *and* the specific questions being posed. With respect to data type, there are some measurement situations, such as determining whether a new blood pressure monitor is sufficiently accurate, where professional organizations or other expert panels set community standards (Asmar & Zanchetti, 2000; JCS Joint Working Group, 2012; Parati et al., 2010, 2014). Whenever available, these guidelines should be adopted. With respect to constraints posed by research questions, these may reflect, for example, the desired use of data in subsequent analyses. For instance, heart rate (HR) is most traditionally derived from an electrocardiogram (ECG) signal, but may also be derived from a photoplethysmographic signal (PPG; the optical HR measure available in most wrist-based devices). Whether a PPG-based measure of HR is sufficiently accurate depends upon the desired use of HR as a dependent variable. For example, if the study goal is to measure high-frequency heart rate variability (HF-HRV; sometimes called respiratory sinus arrhythmia or RSA), then PPG may lack enough temporal precision to detect heartbeats with sufficient fidelity. In this case, ECG-derived HR with a sufficiently high sampling rate is the better measure (Task Force of ESC and NASPE, 1996)[1]. On the other hand, if the goal of a study is to obtain a sufficiently accurate measure of mean HR over larger windows of time that minimizes recording burden on participants, then the accuracy of detecting HR via PPG may be sufficient. In either case, researchers benefit from clarifying their research questions and analysis plans *before* determining their criterion for "sufficient accuracy."

In addition to determining sufficient accuracy, it is important to choose a device with sensors that have sufficient *precision* (i.e., signal-to-noise ratio), *reliability* (i.e., reproducibility), and *feasibility* (given a specific population, purpose, or setting) for the measure of interest. Many of these considerations can affect data quality and therefore should be considered when choosing a device to answer a given research question.

Below we describe a step-by-step framework (see Fig. 1) to guide the selection and assessment of ambulatory physiological devices with respect to these criteria (accuracy, precision, reliability, and feasibility). We then exemplify the use of this framework by reporting on two validation studies conducted by our team. If followed, our framework can help researchers consider various elements of device choice and validation in order to more confidently and reliably answer their own unique research questions.

[1] In fact, we and others have argued it is the only sufficiently accurate measure (see Berntson et al., 1997).


**Fig. 1** Methodological framework. Seven steps for selecting and benchmarking mobile devices in psychophysiological and physical activity research. The arrows indicate that results from steps 6 and 7 can inform the design of additional data collection

**Step 1. Identifying signals of interest** A full discussion on how to determine signals of interest is beyond the scope of this paper; however, we provide some brief guidance in Table 1. In addition, we suggest that readers refer to previously published in-depth resources for additional guidance (Boucsein, 2012; Cacioppo et al., 2017; Stern et al., 2000). Of most importance, signals of interest should be chosen on theoretical grounds or based on prior literature.

**Step 2. Characterizing intended use case** Designing an ambulatory study is often an exercise in balancing the desire for rich data (e.g., longer recording duration, greater number of signals) with higher participant burden and concerns about compliance (e.g., whether participants wear the device(s) as intended). Consequently, considering intended use cases *before* choosing and evaluating specific ambulatory devices can save time and mitigate potential compliance issues.

For each device, researchers should consider user comfort, obtrusiveness, user interface complexity, and data privacy. For example, the comfort of ambulatory devices may depend on body mass index (BMI), sex or gender (e.g., body shape or clothing styles/preferences), daily routines (e.g., exercising, bathing, sleeping), visual acuity, or tactile ability. Bodily location of the device or its sensors could also affect reliability and validity of collected data (for example with skin conductance, see van Dooren et al., 2012). When it is important for sensors to be unobtrusive or not visible to others, researchers should consider devices

**Table 1** Identifying signals of interest commonly measurable via mobile devices

| System | Signal | How to measure |
|---|---|---|
| Electrodermal | Electrodermal activity (EDA) | Two electrodes on hands/fingers, feet/toes, wrist, back, or other sites (van Dooren et al., 2012) |
| Cardiovascular | Electrocardiography (ECG) | Chest strap or a minimum of two electrodes typically on the chest, but perhaps also using the arms, hands, legs, or feet |
| Cardiovascular | Photoplethysmography (PPG) | Optical sensor on the ear, finger, wrist, arm, etc. (Allen, 2007) |
| Cardiovascular | Impedance cardiography (ICG) | Band or spot electrode sensors, typically on chest and back |
| Respiratory | Respiration | Respiration belt on chest or via an impedance-based technique via sensors |
| Physical activity | Accelerometry (and gyroscope) | Accelerometer on waist, chest, arms, head, and/or legs |
| Contextual factors | Location (GPS)<br>Temperature<br>Humidity<br>Audio recordings<br>Video recordings | Smartphone or mobile device: GPS, thermometer, humidity detector, audio recording, etc. |

that can be placed underneath clothing, while being mindful of potential pressure artifacts on sensors. Researchers must also consider whether participants need to access their data themselves (e.g., as is often the case in studies employing biofeedback) and in what situations participants should be offered choice about when they are monitored to mitigate privacy concerns.

Finally, researchers should consider environmental features of the implementation context that can impact device operation or data validity. For example, they should consider environmental features such as electromagnetic interference, changes in ambient lighting, temperature, humidity, altitude, and/or vibration (Strangman et al., 2018; Wilhelm & Grossman, 2010).

**Step 3. Identifying study-specific pragmatic needs** Wearable devices differ in their price, system compatibility, software features (e.g., proprietary vs. compatible with only some operating systems), and battery life. It is important to consider a device's battery life if many signals are recorded, the recording time is long, and/or the sampling frequency is high (Halson et al., 2016). Other device-related features to consider include: form factor (e.g., where the device is worn; Halson et al., 2016); wireless transmission needs (e.g., logging vs. streaming); data storage requirements (e.g., local data storage vs. on a remote server); system functionality (e.g., maximum number of signals that can be recorded); temporal precision (e.g., general trends over longer time periods vs. faster changes at shorter timescales); and dynamic range of the sensors (e.g., large changes in acceleration during sporting events or vehicular travel vs. small changes in acceleration while walking or during other activities of daily living). Finally, it is important to assess whether participants can adequately place sensors on their own body and use devices correctly, including whether they can easily access sensor sites, start and stop recording, and consistently charge devices.

**Step 4. Selecting devices for evaluation** Device options change rapidly, so it is important to identify devices through first-hand experience, recommendations from knowledgeable colleagues, demonstrations at scientific conferences, and searches of the scholarly literature. Some companies offer product demonstrations, which are extremely helpful for interacting with devices first-hand and receiving manufacturer guidance to optimize performance. One must also balance the fact that older devices are sometimes more suitable if they have been used and validated in published research. However, older devices might become obsolete, and may be unavailable for purchase or service/support, or the company that sold them may no longer exist. Another consideration when selecting devices for evaluation are data security protections afforded on the device itself and during transmission of data between the device to or from the cloud or lab servers (e.g., encryption). These features should be selected based on sensitivity of the data being collected and the need for privacy of such data for users.

**Step 5. Establishing an assessment procedure** A validation study should determine the strengths and limitations of different ambulatory devices in contexts similar to those in which they will be implemented (i.e., with similar signals, study populations, implementation contexts either inside or outside the lab). It is also useful to compare device(s) across physical and psychological tasks of varying intensities to test for device sensitivity, floor and ceiling effects of the sensors, and effects of different postures. We recommend selecting well-used and oft-validated tasks wherever possible (for a great example, see Menghini et al., 2019). This enables a researcher to better attribute a validation failure to the specific device being tested, rather than to problems associated with a novel task. Additionally, we highly recommend obtaining qualitative or quantitative user feedback in the form of free-response or survey data. When designing user feedback formats, both open-ended free response and quantified survey responses have unique strengths and weaknesses. Open-ended feedback may unearth unanticipated concerns but can be difficult to interpret. Survey data can be easier to interpret, but requires researchers to successfully anticipate relevant concerns, and also assumes that all individuals utilize survey items identically. In either case, user-feedback data is invaluable for assessing participants' experiences of comfort/discomfort, device obtrusiveness, and the intuitiveness of user interfaces (e.g., ease of starting/stopping recording, putting on and taking off devices either alone or with help) (Nelson et al., 2019; Spagnolli et al., 2014).

**Step 6. Performing qualitative and quantitative analyses on validation data** After pilot data have been collected, we recommend performing a hierarchical set of analyses beginning with the assessment of general trends (for similar method, see Menghini et al., 2019). General trends can be assessed using simple visual inspection of data (e.g., assessing whether a signal increases or decreases as expected). Devices without face validity should not be subject to further testing (e.g., erratic signal, unacceptable signal-to-noise ratio, complete insensitivity to change across conditions expected to elicit change). Once general trends have been established, data quality should be assessed with respect to a gold-standard device. Data quality can be assessed using signal-to-noise ratio, measures of data loss (e.g., missing heart beats), or simple measures of agreement such as Pearson product–moment correlations, intraclass correlations, or Bland–Altman analyses (Bland & Altman, 2007; for example decision criterion see van Lier, et al., 2019). When assessing both general trends and data quality, we endorse recently published guidelines which suggest assessment at the signal level (e.g., raw skin

conductance), the parameter level (e.g., rate of skin conductance responses), and the event level (e.g., rate of skin conductance responses during lower arousal vs. higher arousal scenarios) (van Lier et al., 2019). Finally, qualitative data from user-feedback forms can be explored (e.g., using thematic coding, simple statistics, or visualization) to unearth participant concerns in addition to any individual differences which may have led to usability problems (for examples, see Beukenhorst et al., 2020; Shcherbina et al., 2017).

**Step 7. Conducting power analyses to determine device accuracy** Below, we briefly describe two approaches for conducting power analyses to determine device accuracy (for a more detailed review see Lakens, 2013). In the first approach, a researcher can assess a priori the number of data samples (i.e., instances or individuals) needed to detect significant variation between dependent variables obtained from a new device and from a gold-standard device. This approach requires that researchers determine what they consider to be a meaningful discrepancy in measures *before* conducting their validation study. Critically, what constitutes a "meaningful discrepancy" may differ based on the dependent variable being measured, or that variable's function in subsequent analyses. In the second approach, researchers can *first* conduct a small pilot study, and then use collected data to obtain an effect size estimate. This effect size estimate can then inform how many samples (again, instances or individuals) would be needed to observe a statistically significant difference between two devices for a given power level (often 0.80) and false-positive rate (often 0.05). In both of these approaches, the objective is to enable more rigorous inferences by establishing statistical power *before* collecting independent validation data.

**Illustrative case** Within the bounds of this framework, there are many researcher degrees of freedom. Ultimately, validation studies must be tailored to specific signals, use cases, and research questions. To illustrate how one might use and adapt this framework to a *specific* use case, we describe methods and results obtained in two illustrative studies (detailed in Table 2). Both of these studies were designed to test multiple wearable devices for psychophysiological field experiments in the areas of affective science and health psychology.

## Methods

### Participants

Ten participants from Northeastern University and the surrounding area completed Study 1 (ages 18–36 years, 8 female), and another 11 participants (ages 19–37 years, 1 female) completed Study 2. Per our eligibility criteria, participants were men or women at least 18 years old who were free from significant psychiatric, neurologic, or other medical problems that could place individuals at risk of undue stress, affect their ability to participate fully in the experimental protocol, significantly impact their physiological responses, or adversely impact device testing (i.e., no seizures, head trauma, diagnosed schizophrenia, mood or anxiety disorder, or autism spectrum disorder). Participants provided informed consent under a protocol approved by the Northeastern University Institutional Review Board.

### Procedure

All study procedures were completed in-lab (see Table 2 for rationale). After enrollment and eligibility screening, we measured height, weight, and waist circumference. We then placed the physiological devices shown in Fig. 2 on participants. Next, participants completed a demographic questionnaire (age, race, ethnicity) followed by a 5-min seated rest period. In Study 1, participants completed (in fixed order): (1) a heartbeat detection task (approximately 30 min; Kleckner et al., 2015; Whitehead et al., 1977); (2) an evocative image task (Lang, Bradley, & Cuthbert, 2008); (3) a heartbeat tracking task (Schandry, 1981); (4) a physical activity task (consisting of 30 consecutive squats); and (5) a series of affective and physical activity questionnaires unrelated to the current study. In Study 2, following a 5-minute rest period, participants completed the evocative image task, a physical activity task (30 consecutive squats), and two trials of a mental math task (e.g., Quigley et al., 2002). Each of these tasks was chosen for its demonstrated validity and common usage in affective psychology. In both studies we removed most of the physiological sensors after completion of the experimental tasks and debriefed participants while they completed a final physical activity task (stair climbing). We then removed the remaining sensors and provided $30 remuneration for their time and effort.

### Evocative image task

Participants viewed images (53 pictures in Study 1, 33 in Study 2) from the International Affective Picture System (Lang, Bradley, & Cuthbert, 2008) for 20 minutes. Each trial consisted of a variable 3–8-second "Get Ready" period and a 6-second picture presentation, after which participants rated how pleasant/unpleasant and how activated/deactivated they felt in response to the prior picture. Images with similar normative affect ratings were presented in blocks of 10. Study 1 images were normatively characterized as unpleasant high arousal (e.g., mutilated bodies), unpleasant low arousal (e.g., funerals), pleasant high arousal (e.g., sports), pleasant low arousal (e.g., kittens), and neutral low arousal (e.g., office supplies). Study 2 images included normatively unpleasant high arousal, pleasant high arousal, and neutral low arousal.

**Table 2** Details of workflow across the seven steps of our benchmarking framework applied to two empirical studies presented herein

| | |
|---|---|
| **Step 1. Identify signals of interest** | • Devices should collect data relevant to our primary areas of research: emotion and health. Specifically, devices should be capable of acquiring both heart rate (HR) and electrodermal activity (EDA)—two physiological measures implicated in subjective arousal and psychophysical stress.<br>• Because we are interested in deploying devices in real-world settings, devices should also collect 3-axis accelerometry data (to help account for motion artifacts). |
| **Step 2. Characterize intended use case** | • Devices should be tolerated by healthy young adults during everyday life.<br>• Devices should be unobtrusive, comfortable to wear, and should not limit participant mobility.<br>• Participants should not have to access their data, instrument themselves, or make decisions about when/where they are monitored. |
| **Step 3. Identify pragmatic needs** | • Devices should be financially feasible (less than $500 per unit).<br>• Devices should have sufficient battery life to record several hours (but not necessarily days) of physiological data between charges.<br>• Devices may be located on the wrist, palm, or chest.<br>• We need to assess differences between wet and dry EDA electrodes (i.e., with and without isotonic paste).<br>• Devices should be sensitive enough to detect modest to large changes in physiological activity.<br>• We prefer ability to visualize data in real time to ensure good signal quality prior to deployment.<br>• We have no explicit preferences about data storage. |
| **Step 4. Select devices for evaluation** | • In Study 1, we tested five mobile devices that measured various combinations of EDA, HR, and/or accelerometry from the wrist and chest.<br>• Devices for Study 1 included: Q Sensor device (Affectiva, Boston, MA, USA); the E3 device (Empatica, Milano, Italy); and the Actiwave Cardio device (CamNtech Ltd., Cambridge, UK).<br>• Devices for Study 2 included those tested during Study 1, plus the E4 device (Empatica, Milano, Italy) and the Shimmer EDA device (Shimmer, Dublin, Ireland). |
| **Step 5. Establish assessment procedures** | • We utilized tasks that elicit robust changes in physical activity (squats).<br>• We utilized well-validated tasks that elicit modest to large changes in psychophysiological activity (an evocative image task and a mental arithmetic task), comparable to those we expect to occur during everyday life.<br>• Validation was performed in-lab for comparison against wired gold-standard devices, and to disambiguate device performance from unanticipated context effects. |
| **Step 6. Perform qualitative and quantitative analyses** | • Assessed general trends using visual inspection.<br>• Assessed data quality by estimating signal-to-noise ratio and quantifying data loss.<br>• Assessed qualitative aspects of device performance in addition to user feedback from participant debriefings. |
| **Step 7. Conduct power analyses** | • *Not needed for this work because our primary goals were about general signal trends and not a statistically rigorous comparison across devices.* |

To anchor participants' use of rating scales, the first block of images in each study contained three pictures: one unpleasant high arousal (mutilation), one pleasant high arousal (children on a roller coaster), and one neutral (a basket). One participant's data was excluded from analysis because they reported the images to be too evocative and stopped the picture-viewing task early.

## Physical activity squats task

For the first physical activity task, experimenters guided participants in completing 30 squats followed by two minutes of seated rest.

## Mental math task (Study 2 only)

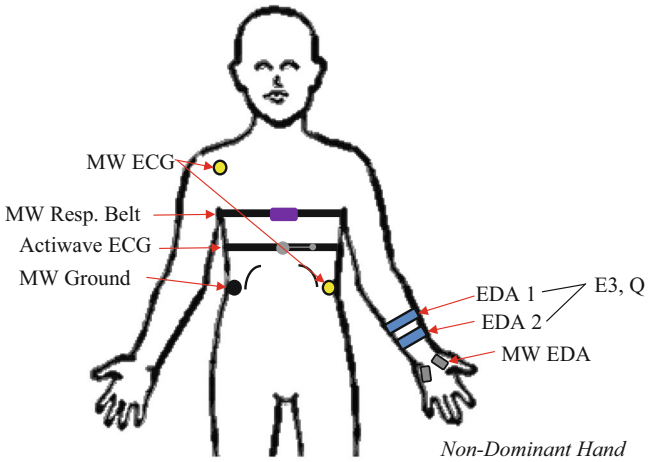
Participants completed two trials of a mental math task (Quigley et al., 2002), wherein participants were instructed to subtract the number 7 from 725 and report their answers aloud as quickly and as accurately as possible. Serial subtractions were supervised by an experimenter trained to provide feedback ("incorrect") whenever the participant provided an incorrect answer. Following feedback, the experimenter prompted the participant to resume subtractions from the last correct response. The first trial lasted 60 seconds, after which the experimenter left the room and a 2-minute resting baseline was recorded. Trial 2 of mental math was identical to trial 1, except the difficulty level of the second trial was determined based on the participant's performance during the first trial. If the participant answered fewer than five responses correctly in trial 1, then trial 2 was made easier (subtracting 6 from 847); otherwise the second trial was made harder (subtracting 13 from 847). Trial 2 was followed by a second 2-minute baseline.

## Ambulatory devices

In Study 1, we tested three ambulatory devices that each measured one or more of the following: EDA, HR, and/or

**STUDY 1**

1. Demographic questionnaire
2. 5-min seated rest
3. Heartbeat detection task
4. *** Evocative image task**
5. Heartbeat tracking task
6. Questionnaires
7. *** Squats**
8. Stair climb & debrief




**STUDY 2**

1. Demographic questionnaire
2. 5-min seated rest
3. *** Evocative image task**
4. *** Squats**
5. *** Mental math**
6. Stair climb & debrief

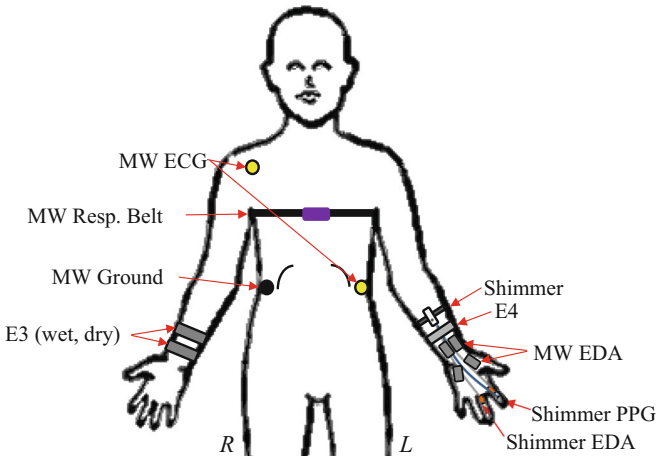



**Fig. 2** Study flow and device placements for Studies 1 and 2. In Study 1, the position of the EDA 1 and EDA 2 devices (E3 and Q devices) were counterbalanced across participants. We only present data from tasks that are bolded and starred (*)

accelerometry from the wrist or chest (Table 3). These included the Q Sensor device (Affective, Boston, MA, USA; Poh et al., 2010), E3 (Empatica, Milan, Italy), and Actiwave Cardio (CamNtech Ltd., Cambridge, UK). Data

**Table 3** Devices used in Study 1. Hz = Samples per second for data acquisition; EDA = electrodermal activity; BVP = blood volume pulse; ECG = electrocardiogram

| Device | Location | EDA Sensor | Heart Rate | Accelerometry |
|---|---|---|---|---|
| 1. Affectiva Q Sensor | Non-Dominant Wrist | Dry; 32 Hz | N/A | 32 Hz |
| 2. Empatica E3 | Non-Dominant Wrist | Dry; 4 Hz | BVP; 64 Hz | 32 Hz |
| 3. MindWare EDA | Non-Dominant Palm | Wet; 1000 Hz | N/A | N/A |
| 4. Actiwave Cardio | Chest | N/A | ECG; 1024 Hz | 32 Hz |
| 5. MindWare ECG | Chest | N/A | ECG; 1000 Hz | No |

from these devices were compared to a research-grade wired laboratory system from MindWare Technologies, Ltd. (Gahanna, OH, USA), which served as our gold standard and which sampled ECG and EDA at 1000 Hz using BioLab v. 3.0.13 software (MindWare Technologies, Ltd.) and a BioNex 8-Slot Chassis (model 50-3711-08). In Study 2, we focused specifically on EDA measurements by comparing five different device configurations (Table 4). First, we tested for differences in electrode type using dry electrodes (no isotonic paste) vs. wet electrodes (with isotonic paste) with the E3 on the wrist. Second, we tested for differences in recording location comparing wrist vs. the palm. Third, we compared two additional devices not tested in Study 1, namely, E4 (Empatica, Milan, Italy) and Shimmer EDA devices (Shimmer, Dublin, Ireland). For sampling rates, see Tables 3 and 4. The Q Sensor had been previously used, so its durability may have been affected by prior use.

Before testing, each device was time-synchronized with the computer used to acquire data from the MindWare system. The Shimmer device inexplicably did not synchronize its clock to the computer's clock despite following manufacturer instructions and thus was time-synchronized to the E4 device's clock within 100 milliseconds by manually aligning accelerometer data in the physical activity task for each participant. This approach for synchronization should minimally influence results. All of the ambulatory devices recorded data to internal memory. Although many of the ambulatory devices tested provided adequate ways to visualize data pre- and post-acquisition, some did not, which made it difficult to anticipate future recording issues during acquisition (e.g., loose sensors). All devices were used per manufacturer's recommendations, and all sensors were worn for at least 10 minutes before recording the data presented in this manuscript.

## Data analysis

Data from each device were downloaded and processed as suggested by the manufacturer. For EDA data, we distinguished between tonic, background skin conductance level (SCL) trends, and rapid phasic skin conductance responses (SCRs), and we focused our analysis on SCL and rate of SCRs. ECG and blood volume pulse (BVP) were analyzed to obtain inter-beat intervals (IBIs) using MindWare's HRV analysis program, and subsequently all results were visually inspected for general trends.

To determine signal-to-noise ratios for EDA data, we first calculated the magnitude of the signal as the maximum SCL minus the minimum SCL during physical activity. Next, we quantified noise as the standard deviation of the EDA signal in a relatively stable 12-second segment of data where no SCRs were evident. After selecting this 12-second segment for each participant and device, we removed slow trends in SCL by subtracting the linear best-fit line from the 12-second segment of data. We ignored data from participants with loose sensors where signal quality was extremely poor, as these records do not reflect the true capabilities of the devices under study. Our strategy for calculating HR signal-to-noise ratios was identical to that used for EDA data, except the duration of the segment used to compute noise was 1 second instead of 12 seconds. A 1-second duration was chosen between heartbeats when the ECG signal was near its isoelectric line. These analyses and visualizations utilized in-house software programmed in MATLAB (MathWorks, Natick, MA, USA).

**Table 4** Devices used in Study 2. Hz = Samples per second for data acquisition; EDA = electrodermal activity

| Device | Location | EDA Sensor | Sampling Rate |
|---|---|---|---|
| 1a. Empatica E3 | Right Wrist | Dry | 4 Hz |
| 1b. Empatica E3 | Right wrist | Wet | 4 Hz |
| 2. Empatica E4 | Left Wrist | Dry | 4 Hz |
| 3. Shimmer | Left Fingers | Dry | 51.2 Hz |
| 4a. MindWare EDA | Left Palm | Wet | 1000 Hz |
| 4b. MindWare EDA | Left Wrist | Wet | 1000 Hz |

## Results

### Overview

We assess general trends and data quality for EDA data from all devices during the physical activity task, the evocative images task, and the mental math task. Next, given our experimental goals, we compare HR and accelerometry data across devices during the physical activity task. Data from other task/device combinations are beyond the scope of this paper. Given the small sample size, we often illustrate data for individual participants. Qualitative comparisons and inferences are included in the discussion section.

### Electrodermal activity during physical activity

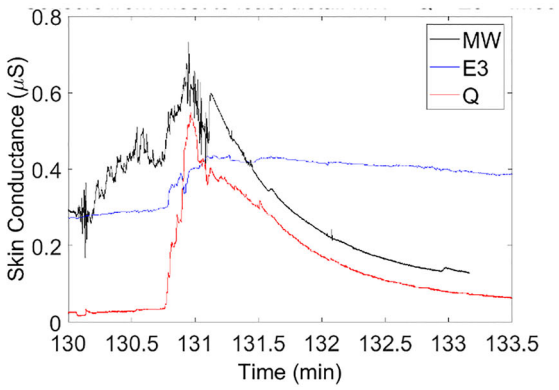
We expected SCL and rate of SCRs/minute to increase during squats, and then to decrease during the subsequent two minutes of seated rest. In Study 1, the gold-standard in-lab MindWare EDA analysis software revealed the expected trend in 7 of the 10 participants from Study 1. Of the remaining three participants, one (#3) appeared to be electrodermally stabile with a virtually unchanging SCL, one (#4) exhibited many SCRs but no change in SCL, and one (#8) had poor data quality due to a poor electrode connection. By comparison, although the ambulatory EDA devices often showed the expected pattern of SCL increase during physical activity followed by a decrease during seated rest, they typically evidenced a smaller dynamic range of SCL and many fewer SCRs (Fig. 3). Specifically, the E3 showed the expected increase in SCL following the onset of squats in 7 of 10 participants, and 5 of those 7 participants showed an expected decrease in SCL during subsequent rest. Additionally, the E3 was the only ambulatory EDA monitor to detect some SCRs during squats (in six of nine participants who showed multiple SCRs as measured by the gold-standard MindWare device). None of the other ambulatory devices detected SCR counts/minute that approached the number detected by the MindWare device. The Q Sensor device showed an expected increase in SCL following onset of squats in three of eight participants, and an expected decrease in SCL during subsequent rest for two of eight participants. MindWare had the highest signal-to-noise ratio (550 ± 506), followed by the E3 (525 ± 660), and finally the Q Sensor (217 ± 359; Fig. 4).

In Study 2, we compared data from dry vs. wet EDA electrodes using the wrist-based EDA devices. The E3 with wet sensors performed best, showing the greatest changes in SCL, even larger than the gold-standard MindWare device for some participants. The dry E4 performed least well, showing the smallest changes in SCL. We then compared wrist-based to palm- and finger-based placements across devices. Physical activity resulted in greater SCL changes from some of the wrist-based placements (E3 dry, E3 wet, MindWare wet) when compared to palm- and finger-based placements (MindWare palm, Shimmer finger; Fig. 5). In contrast, palm- and finger-based placements showed a higher rate of SCRs/minute than wrist-based placements.

### Electrodermal activity during evocative images

In Study 1, MindWare EDA data revealed a high rate of SCRs—many of which were large—for three of nine participants, a modest rate of SCRs for three participants, and virtually no SCRs for three more participants (Fig. 6). All three ambulatory EDA devices failed to detect most SCRs evident from the MindWare device during the image viewing task in both highly reactive individuals (participants 4, 6, and 8) and modestly reactive individuals (participants 1, 2, and 5). Because devices did not achieve face validity, we did not proceed with subsequent analysis.

Data from Study 2 revealed better device performance with palm- and finger-based placements (MindWare wet electrodes on the palm, Shimmer dry electrodes on the fingers) compared to wrist-based placements (MindWare wet, E3 wet, E3 dry,



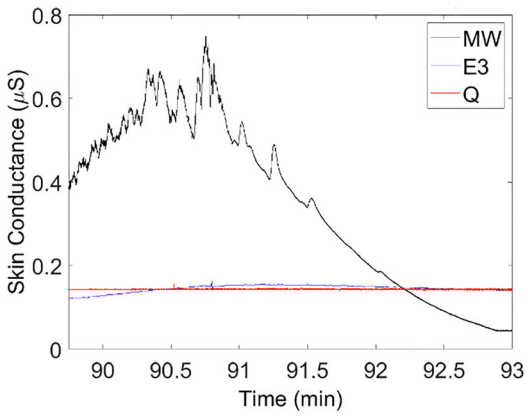




**Fig. 3** EDA data during physical activity squats task. **Left**: Example showing the highest correspondence between the MindWare (MW) EDA device and mobile EDA devices (participant #7). **Right**: Example showing the lowest correspondence between the MindWare EDA device and mobile EDA devices (participant #9)

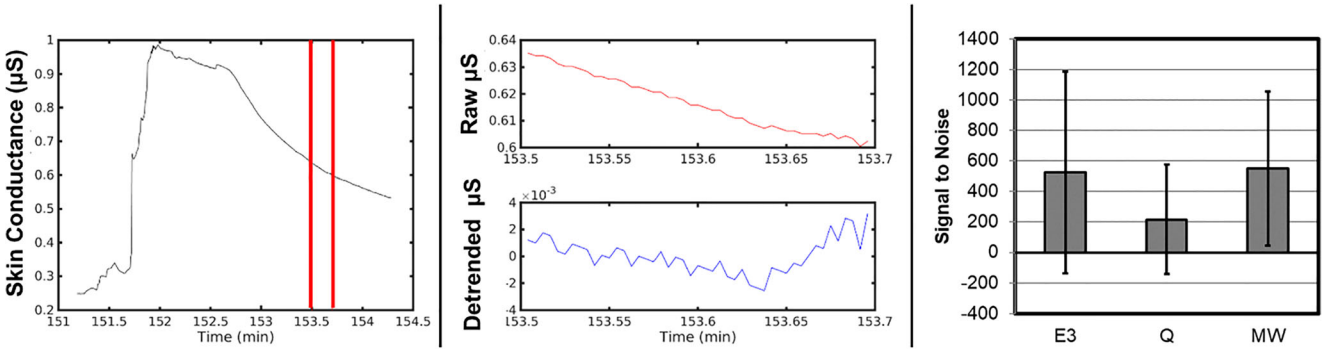



**Fig. 4** Signal-to-noise ratio in EDA data during physical activity. **Left**: EDA data for one participant (#6) and device (E3) during 3 minutes of physical activity. The two vertical red lines starting at min 153.5 indicate the 12-second segment used to compute the signal-to-noise ratio. **Center**: Top plot shows raw EDA data in the same 12-second segment. The bottom plot shows linearly de-trended data in the 12-second segment, where standard deviation was used as a measure of noise. **Right:** Average and standard deviation in signal-to-noise ratio across participants for each device

and E4 dry). This is consistent with results from wrist-based placements in Study 1, which performed more poorly than palm-based placements using wet sensors with the MindWare device. Figure 7 shows representative samples of data from two participants. In line with prior research, palm- and finger-based placements better detected SCRs (as demonstrated in both studies) likely due to greater concentration of eccrine sweat glands on the palmar surface of the hand than on the wrist (Boucsein, 2012). Due to poor measurement of wrist-measured SCRs during evocative images in Study 1, we introduced an additional task in Study 2 (mental math task) to induce greater electrodermal activity and thereby better distinguish among devices by avoiding floor effects.

## Electrodermal activity during mental math

As expected from prior work, the mental math task induced measurable SCRs in more participants (8 of 11 participants) than the evocative images task (5 of 11 participants in Study 2). Further, the mental math task led to some measurable SCRs from the wrists of some devices for several participants. Nevertheless, consistent with the evocative images task, palm- and finger-based placements better detected SCRs during the mental math task (MindWare on palm, Shimmer on fingers) than did wrist-based placements (MindWare on wrist, E3 wet, E3 dry, and E4 dry; see Fig. 8).

## Heart rate during physical activity

The Actiwave ECG device performed well compared to our gold-standard MindWare ECG device (Study 1 only). Figure 9 illustrates that high-quality data were routinely observed from the heart rate devices when participants were stationary. However, when participants were squatting, data from the heart rate devices exhibited substantial movement artifacts when the signal was near the isoelectric line, although R-spikes were still apparent in both the Actiwave and MindWare ECG-based data (Fig. 9, right).

During physical activity, the ambulatory Actiwave ECG device outperformed even the gold-standard MindWare

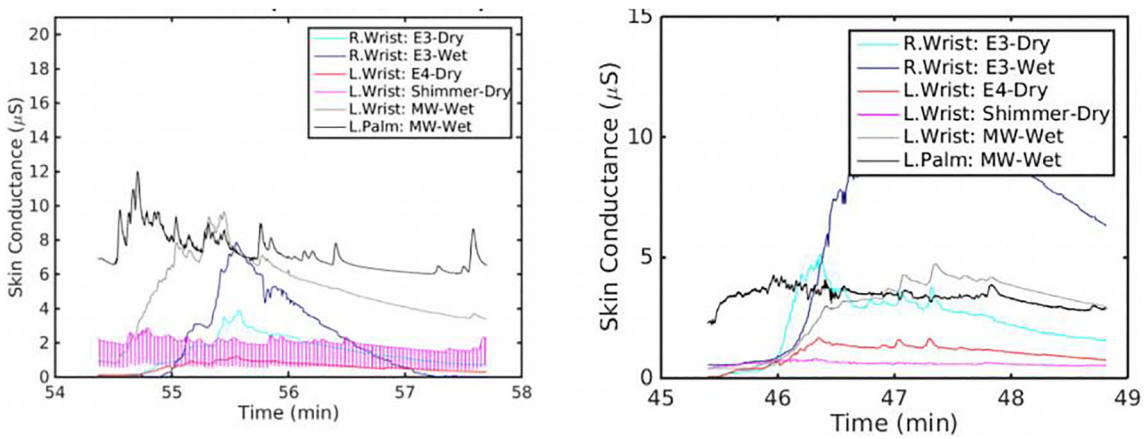



**Fig. 5** EDA data during physical activity from Study 2 participant 14 (left panel) and participant 19 (right panel) which illustrates that the wrist-based device placement evidenced greater changes in SCL compared to devices using palm- and finger-based placements. However, palm- and finger-based placements showed greater SCR rates/minute.

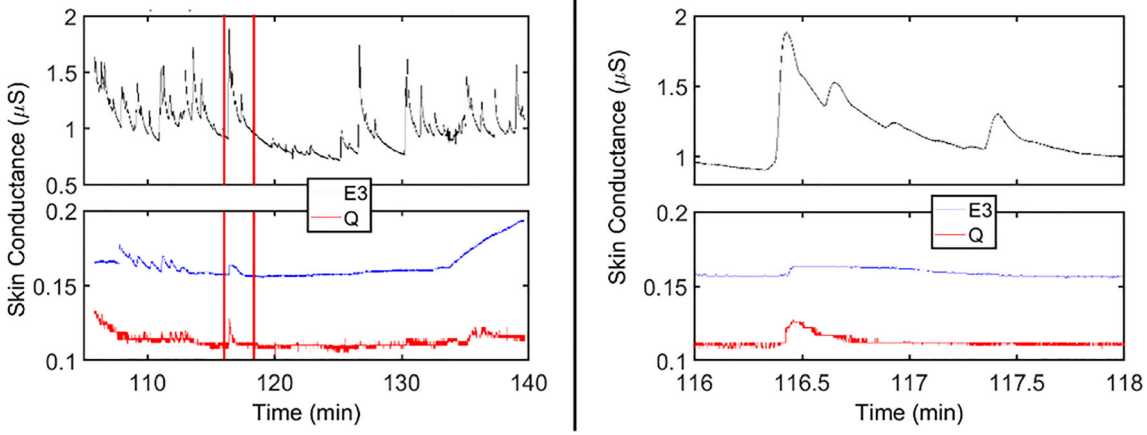



**Fig. 6** EDA data from participant 6 during the evocative image task. **Left:** These panels depict data from the participant who demonstrated the largest SCRs using the E3 and Q sensor during evocative picture viewing. Amplitudes of SCRs measured by the lab-based MindWare EDA device (top panels) are much larger and reveal many more SCRs than observed with ambulatory EDA devices (bottom panels). **Right:** The right panels show a zoomed-in view of a portion of data showing the largest amplitude SCR from the left panel between the two vertical red lines from minutes 116 to 118

ECG device, presumably because the Actiwave was affixed to the chest, whereas the MindWare device has long wires that can result in motion-related artifacts. By comparison, the E3 BVP did not perform well either during movement or when participants were stationary; specifically, 5 of 10 participants exhibited significant artifacts that precluded analysis of HR data from the E3 BVP. This is not surprising given that the E3 relies on an optically derived BVP signal that is much more affected by movement artifacts than an electrical signal (i.e., ECG) collected via wet electrodes affixed to the skin.

Quantitative IBI analyses corroborated our visual inspection of raw data: Actiwave generally outperformed the lab-based MindWare ECG device when a participant was engaged in repetitive squats. Figure 10 shows that during squats, MindWare and E3 devices failed to detect some R-spikes in the ECG. However, when participants were still, IBI results agreed well across MindWare and Actiwave devices, and, to a lesser extent, with the E3.

Quantitatively, our results in Table 5 show the fraction of heartbeats that were not detected by each device for each participant. We calculated this for each participant by comparing the observed number of heartbeats detected by each device to the maximum number of heartbeats observed across all devices. Our results show that the Actiwave performed best (missing 3 ± 5% of heartbeats), followed by MindWare (missing 6 ± 10% of heartbeats), and lastly the E3 (missing 15 ± 15% of heartbeats) across all heartbeats for all participants during the task. Finally, using data from a sedentary period, MindWare exhibited the highest signal-to-noise ratio (322 ± 250), followed by Actiwave (171 ± 57) and E3 (157 ± 101; Fig. 11).

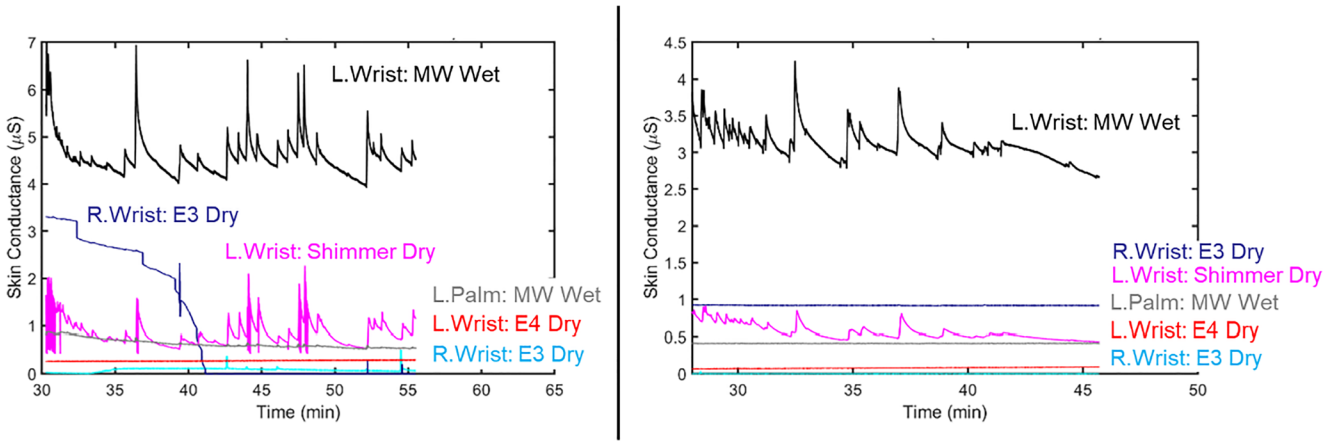



**Fig. 7** EDA data from Study 2 participants 17 and 21 (left and right, respectively) during the evocative images task illustrates that wet sensors on the palm using the MindWare device performed best, followed by Shimmer dry sensors on the fingers. In contrast, we observed poor performance from wrist-based placements on all devices (E3 dry, E3 wet, E4 dry, and MindWare wet)

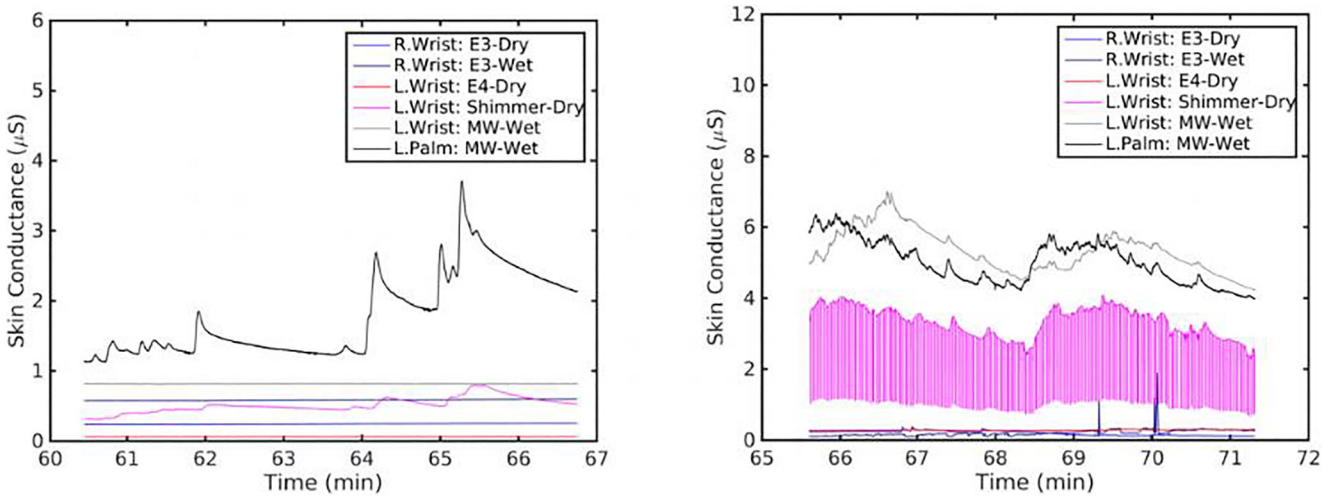



**Fig. 8** EDA data during mental math task for Study 2 participants 12 and 17 (left and right panels, respectively) generally revealed superior performance from palm- and finger-based placements (MindWare on palm, Shimmer on fingers) compared to wrist-based placements

### Heart rate during evocative images

The heart rate data recorded during the evocative picture task was generally of high quality for the mobile devices, as expected, because the participants were stationary during the task. We do not show these data because they do not reveal any substantial differences in performance across the devices.

### Accelerometry during physical activity

We used squats as a benchmarking task to compare mobile accelerometers as we had no gold-standard accelerometer. Figure 12 shows that data from all accelerometers appeared to work well in capturing the squatting motion, as each squat can be seen individually in the accelerometry data.

## Discussion

We describe a systematic benchmarking framework for selecting, testing, comparing, and documenting differences among commercially available, wearable physiological and physical activity devices. We demonstrate use of this framework in two intensive small-sample studies that compared 15 device configurations for 5 EDA devices, 3 heart rate devices, and 3 accelerometers across laboratory tasks designed to elicit either physical activity or subjective and physiological arousal (Tables 3 and 4). Per our framework, we used qualitative *and* quantitative metrics to judge device performance. Moreover, using data from Study 1, which suggested floor effects in the evocative images task, we instituted another evocative task, mental arithmetic, to avoid both floor and ceiling effects in Study 2. Next, we discuss our impressions of each device and of the signals they produced to illustrate how one might reach

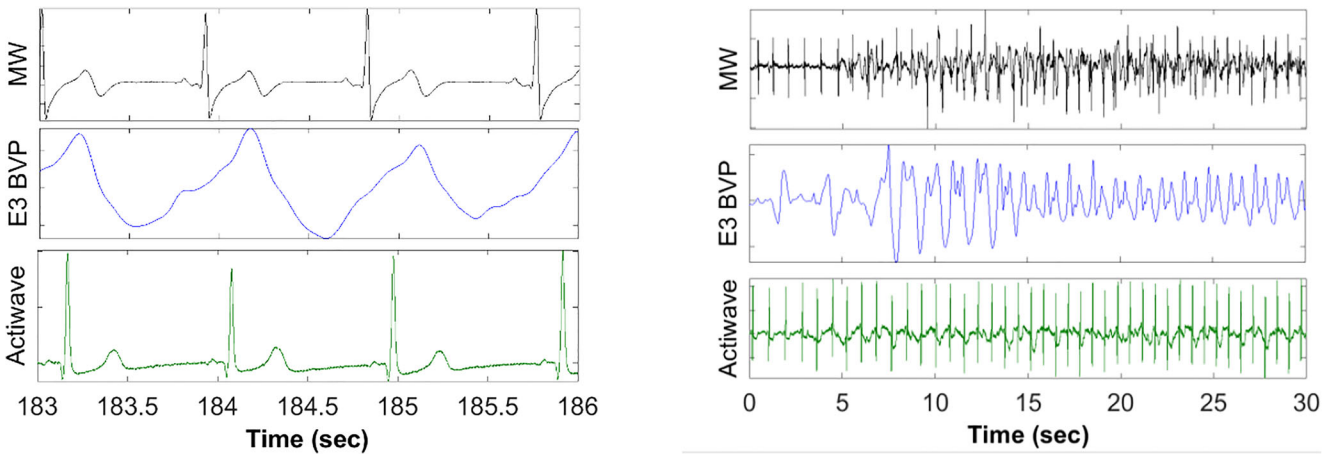



**Fig. 9** Data from HR devices in Study 1. **Left:** Data from participant 6 while stationary. Data are synchronized in time only to within 1 second, so heartbeats do not perfectly align in time across devices. **Right:** Data from participant 3 while performing squats. MindWare ECG and E3 BVP were particularly affected by participant motion, whereas Actiwave ECG data exhibited minimal motion artifacts

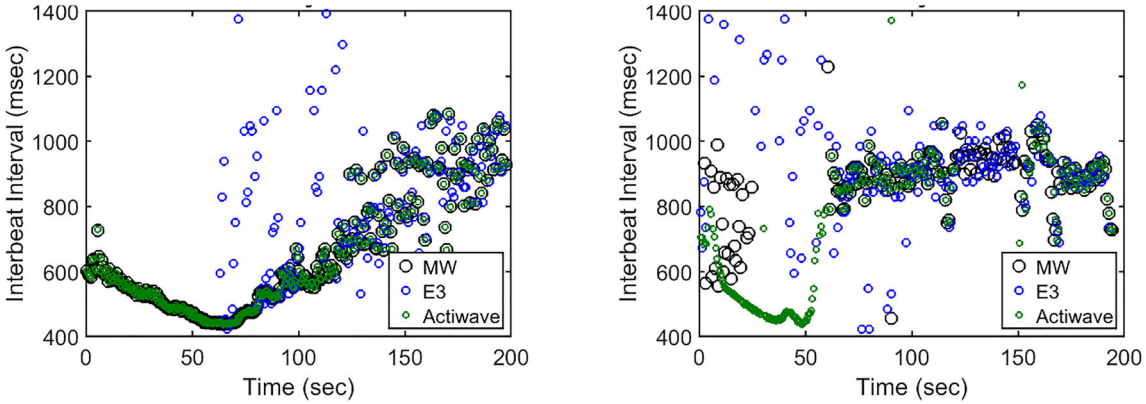



**Fig. 10** IBI data across devices during physical activity in Study 1. **Left:** Example showing good correspondence between devices, especially MindWare and Actiwave (participant 1). **Right:** Example showing the highest quality IBI data for Actiwave, with less stable detection of R-spikes from the ECG data during squats from both the MindWare and E3 devices (participant 5). The participant was repetitively squatting during the period of approximately 0–30 seconds on the *x*-axis (time)

conclusions about which device(s) to choose for research use in the context of our suggested framework.

## Impressions of EDA signals

We observed four key themes regarding EDA data that could help other researchers when considering what device(s) to use in their studies. First, wet EDA electrodes yielded greater changes in SCL compared to dry electrodes, likely increasing sensitivity to change. The use of electrodes with paste is standard in laboratory-based EDA measurement (Boucsein et al., 2012). Second, finger- and palm-based measures were consistently better than those taken from the wrist, corroborating standard recommendations to record EDA activity from the volar (inside) surface of the hands (or feet; Scerbo, Freedman, Raine, Dawson, & Venables, 1992; Venables & Christie, 1980) as well as with more recent comparisons of hand and wrist placement sites (van Dooren et al., 2012). Specifically, we observed more SCRs from the gold-standard MindWare wet sensors on the palm, followed closely by Shimmer dry sensors on the fingers compared to wrist placements with other devices, which were inadequate to detect SCRs. In general, dry sensors are expected to provide smaller and noisier signals because sensors may not consistently cover a specific patch of skin with a given set of eccrine glands, and may slide relative to the skin, creating movement artifacts. We also observed greater changes in SCL from wrist than hand placements, where we observed roughly equivalent performance between MindWare wet sensors on the wrist and E3 wet sensors on the wrist, followed closely by E3 dry sensors on the wrist. Third, the mobile EDA devices produced more false negatives than false positives in that the number of SCRs seen using the mobile devices were a subset of the number of SCRs seen when using the gold-standard MindWare EDA device. Finally, the squats task involving physical activity induced greater changes in SCL, whereas the tasks involving greater

**Table 5** Fraction of heartbeats that were not detected in analysis of HR data during physical activity. Lower percentages (whiter cells) of missed heartbeats reflect higher-quality results, whereas higher percentages (redder cells) reflect lower-quality results (fewer detected heartbeats)

| Participant | MindWare | E3 | Actiwave |
|---|---|---|---|
| 1 | 0% | 48% | 0% |
| 2 | 18% | 23% | 2% |
| 3 | 6% | 4% | 11% |
| 4 | 0% | 13% | 2% |
| 5 | 29% | 25% | 11% |
| 6 | 1% | 7% | 0% |
| 7 | 0% | 7% | 1% |
| 8 | * | * | * |
| 9 | 0% | 1% | ** |
| 10 | 3% | 12% | 0% |
| **Mean and St. Dev.** | 6 ± 10% | 15 ± 15% | 3 ± 5% |

*Data from participant 8 are not shown due to poor data quality from a poor electrode connection

**Actiwave data from participant 9 are not shown because the data were lost

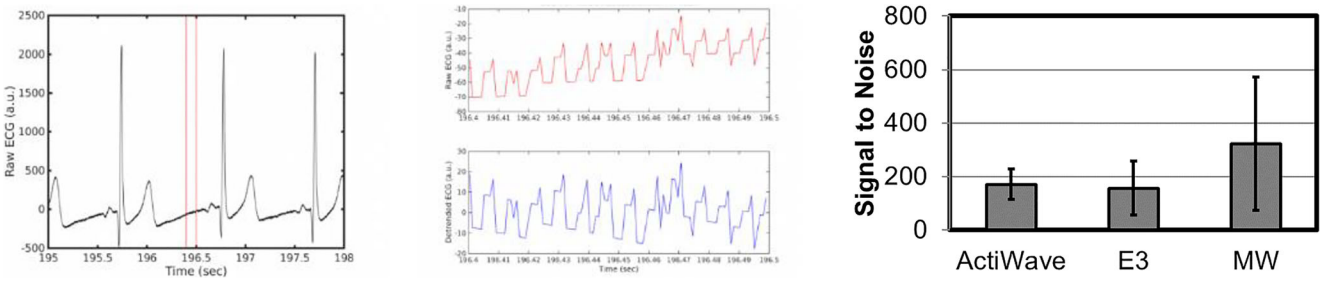



**Fig. 11** Signal-to-noise ratio in HR data while a participant is stationary before physical activity. **Left:** ECG data for one participant (#1) and device (Actiwave) during physical activity. The two vertical red lines around 196.5 second indicate the 1-second segment used to compute the noise. **Center:** Top plot shows raw ECG data in the 1-second segment. Bottom plot shows linearly detrended data in the 1-second segment, where standard deviation was used as a measure of noise. **Right:** Average and standard deviation in signal-to-noise across participants within each device

subjective arousal (based on prior literature with these tasks) induced greater changes in SCR rate.

## Impressions of heart rate signals

The mobile HR devices matched the performance of a gold-standard HR device when participants were stationary, and one device, the Actiwave, exceeded the performance of the gold-standard device when participants were moving. The Actiwave ECG device is securely fixed to the torso using a chest strap, unlike the MindWare lab-based device, which has wires that can move relative to the sensor during participant movement. However, one problem with the Actiwave is that its data quality cannot be assessed during recording. Thus, we recommend making a short recording to first verify data integrity before initiating a longer recording in the field. The E3 BVP was the only device with an optical HR sensor, and it performed less well than the devices that recorded an ECG. Under optimal conditions (i.e., when participants were motionless), E3 BVP data matched that of other devices, but its performance suffered with even small amounts of movement.

## Impressions of accelerometer signals

All accelerometer devices performed well when participants were moving (during repetitive squats).

## Impressions of device construction and usability

The E3 appeared to be durable and well-constructed. For devices that permitted it, data viewing was easy both pre- and post-acquisition using a Mac, iPad, or iPhone. However, the

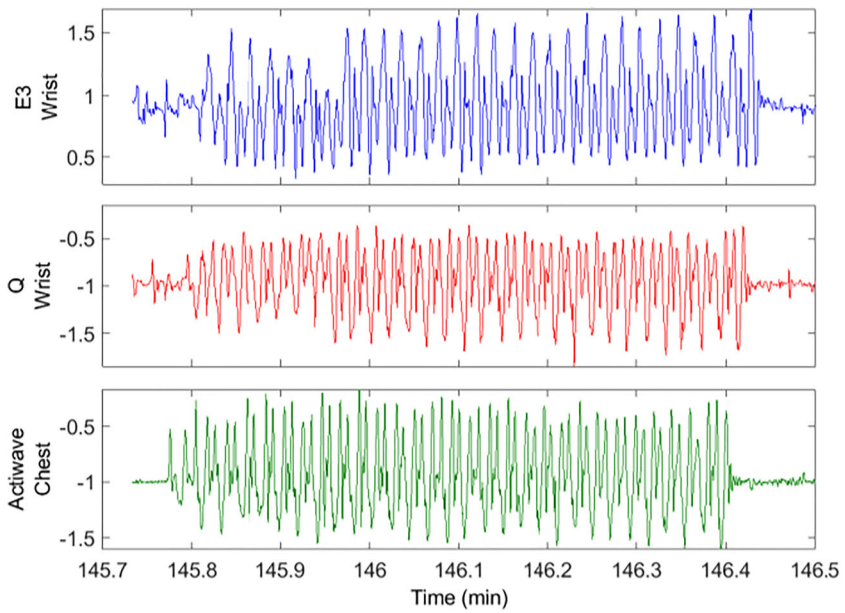



**Fig. 12** Accelerometry data for all three ambulatory devices from Study 1 participant 3 during the squatting portion of the physical activity task, preceding seated rest. Data are shown from the axis (*x*, *y*, or *z*) that best captured the squatting motion for each device. These data are representative of the accelerometry data for all remaining participants

Velcro wristband of the E3 sometimes loosened, so we secured it using medical tape. Furthermore, some participants reported discomfort, suggesting questionable suitability for longer-term recordings. Using wet sensors with the E3 was difficult because electrodes often disconnected or got stuck together, causing data loss. The Q Sensor was reported to be most comfortable due to its durable elastic band. The Actiwave Cardio seemed sturdy, and participants reported it to be comfortable, unobtrusive, and discrete as it was hidden beneath clothing. The E4, like the E3, also appeared to be durable and well-constructed. Both the E3 and E4 had simple, user-friendly software for setup, data download, and data viewing. The E4 was easy to turn on, LED signals were intuitive, and to our research team was the most aesthetically pleasing. However, it was difficult for participants to place the E4 on themselves. The Shimmer EDA had the most complete and intuitive computer interface (ConsenSys program), which displayed many options for data collection. The device readily accepts wet EDA electrodes. However, it was difficult to wrap the device around the participants' fingers, the start/stop button was hard to reach, and the LED signals were not intuitive. Furthermore, the Shimmer sensor housing was not as durable as the Q Sensor and Empatica (E3 and E4) devices, which should be taken into consideration for longer-term deployments.

## Strengths of our benchmarking framework

Our suggested framework provides guidance for selecting and comparatively evaluating ambulatory peripheral physiological and physical activity devices. Our small-scale performance studies have several strengths, including the use of multiple devices and device configurations (e.g., wet vs. dry EDA electrodes). Specifically, we compared 15 device configurations across 5 EDA devices, 3 heart rate devices, and 3 accelerometers. We utilized multiple, well-established laboratory tasks, including those involving either physical activity or psychophysical arousal. We used tasks that were sufficiently activating to distinguish performance across devices across tasks. We also used gold-standard devices as comparators for EDA and heart rate devices. Gold-standard comparators are invaluable for establishing strong validity. Often, users of wearable devices do not expect strong agreement with gold-standard devices; however, *even in this case* it is important to know to what extent a wearable does or does not deliver on this expectation. As we illustrate, many devices are in fact comparable to or, in some cases, *better* suited for a given situation than their gold-standard counterparts (e.g., a chest strap for ECG was better at preventing movement artifacts than wired electrode-style ECG). Finally, we used both qualitative and quantitative comparisons across devices with multiple assessments for data quality, usability, and user comfort.

Our suggested benchmarking framework compliments and extends other frameworks (e.g., van Lier et al., 2019) and validation studies (e.g., Kasos et al., 2019; van Lier et al., 2019) for assessing mobile devices by including considerations for the broader set of decisions that researchers must make prior to initiating a study. That is, our framework emphasizes *selection* of mobile devices based on signals of interest, intended use cases, pragmatic needs (steps 1–4), and establishes an effective assessment procedure to test the strengths and limitations of selected devices (step 5). These initial steps are critical to include because they emphasize the fact that validity is established (or not) only for a particular context (e.g., setting, sample, recording interval) and does not necessarily generalize beyond that context.

## Study limitations

A limitation of our performance studies is their small sample sizes ($N = 10$ in Study 1 and $N = 11$ in Study 2), which reduce the possible range of variability in our results when comparing across devices. However, validation studies are often small because their purpose is to make a rapid assessment of device performance with minimal time and resources invested. In addition, each participant wore multiple devices, allowing for within-person comparisons, thereby increasing sensitivity to across-device differences. Further, we recorded enough data from each participant (more than 35 minutes) and across enough conditions to make a good assessment of data quality. Indeed, when devices produced poor quality data, it was generally evident even with relatively minimal data. Another limitation in our assessments is that differences across devices could be due to varying filter settings or other data acquisition features (e.g., sampling rates), some of which were not made available by the device manufacturers. Such differences can make it difficult, if not impossible, to design identical comparisons among devices.

Given the above caveats, we urge readers not to rely on results obtained in our small-scale studies when choosing specific devices, as reported data are provided only to illustrate our benchmarking framework and the types of conclusions it enables researchers to make. Indeed, some devices (e.g., E3, Q Sensor) are no longer available for purchase, and we make no endorsement regarding any of the devices evaluated herein. Instead, we suggest that researchers assess devices that meet their own pragmatic, research, and data needs. Further, researchers should test devices under the experimental conditions and with the kinds of participants that they wish to include in their own studies.

## Conclusions

We present a benchmarking framework for designing and conducting comparative evaluation studies of wearable

physiological and physical activity devices that we hope will serve as a complimentary addition to published validation procedures in the scientific literature. In particular, our framework aims to be both multi-level and multi-purpose. While there is no one-size-fits-all approach when it comes to empirically validating devices for particular research questions and contexts, we highlight strategies and methods that may be generally applied. Our two small-scale studies illustrate the merits of this framework. Finally, in an effort to increase transparency and rigor in future scientific studies, we encourage authors to provide evaluative, validation data as we have done here either in supplemental materials or in published reports when using consumer-grade or other wearable devices that have insufficient, publicly available evidence of data quality. In particular, we advocate for the inclusion of both quantitative and qualitative user feedback, and remind readers that validation is not a one-time process. Rather, a validation assessment (for a device or measure) is best considered as an ongoing, iterative process performed in a speficic context and for some specific purpose.

**Acknowledgements** The authors thank the volunteer research assistants of the Interdisciplinary Affective Science Lab who helped with data acquisition and analysis of reported data, including Rija Ashfaq, Aileen Gabriel, Anna Neumann, Anthony Siena, and Vishal Sharoff. We thank Dr. Amber Kleckner for her editorial review of our manuscript.

**Funding/Support** This work was supported by a grant to MSG and a subaward to KQ from Janssen Research and Development. IRK was supported by the Janssen Research grant to MSG, NIMH F32MH096533, and NCI grants K07CA221931 and UG1CA189961. Salary support for this work was also provided by the Army Research Institute (W911NF-16-1-0191) and the NIH grants (1U01CA193632-01A1 and R01MH113234).

## Compliance with ethical standards

**Conflict of interest** MSG and KQ have received research and consulting funding from Janssen Research & Development, LLC. MSG also serves on the Scientific Advisory Board at Affectiva, Inc. (Q Sensor manufacturer) and Empatica, Inc. (E3 and E4 manufacturer). All other authors declare no potential conflicts of interest.

**Open practices statement** All materials and data available upon request. None of the experiments described were pre-registered.

## References

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, *28*(3), R1–R39. https://doi.org/10.1088/0967-3334/28/3/R01

Asmar, R., & Zanchetti, A. (2000). Guidelines for the use of self-blood pressure monitoring: A summary report of the First International Consensus Conference. Groupe Evaluation & Measure of the French Society of Hypertension. *Journal of Hypertension*, *18*(5), 493–508.

Berntson, G. G., Bigger, J. T., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., … van der Molen, M. W. (1997). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology*, *34*(6), 623–648.

Beukenhorst, A. L., Howells, K., Cook, L., McBeth, J., O'Neill, T. W., Parkes, M. J., Sanders, C., Sergeant, J. C., Weihrich, K. S., & Dixon, W. G. (2020). Engagement and participant experiences with consumer smartwatches for health research: Longitudinal, observational feasibility study. *JMIR MHealth and UHealth*, *8*(1), e14368. https://doi.org/10.2196/14368

Bland, J. M., & Altman, D. G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics*, *17*(4), 571–582. https://doi.org/10.1080/10543400701329422

Boucsein, W. (2012). *Electrodermal activity, 2nd ed*. https://doi.org/10.1007/978-1-4614-1126-0

Cacioppo, J. T., Louis, T. G., & Berntson, G. G. (Eds.). (2017). *Handbook of Psychophysiology* (4th). Cambridge University Press.

Conner, T. S., Tennen, H., Fleeson, W., & Barrett, L. F. (2009). Experience sampling methods: A modern idiographic approach to personality research. *Social and Personality Psychology Compass*, *3*(3), 292–313. https://doi.org/10.1111/j.1751-9004.2009.00170.x

Ebner-Priemer, U. W., Kuo, J., Schlotz, W., Kleindienst, N., Rosenthal, M. Z., Detterer, L., … Bohus, M. (2008). Distress and affective dysregulation in patients with borderline personality disorder: A psychophysiological ambulatory monitoring study. *The Journal of Nervous and Mental Disease*, *196*(4), 314–320. https://doi.org/10.1097/NMD.0b013e31816a493f

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Gao, W., Brooks, G. A., & Klonoff, D. C. (2018). Wearable physiological systems and technologies for metabolic monitoring. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, *124*(3), 548–556. https://doi.org/10.1152/japplphysiol.00407.2017

Goodwin, M. S., Velicer, W. F., & Intille, S. S. (2008). Telemetric monitoring in the behavior sciences. *Behavior Research Methods*, *40*(1), 328–341. https://doi.org/10.3758/BRM.40.1.328

Goodwin, M. S., Mazefsky, C. A., Ioannidis, S., Erdogmus, D., & Siegel, M. (2019). Predicting aggression to others in youth with autism using a wearable biosensor. *Autism Research: Official Journal of the International Society for Autism Research*, *12*(8), 1286–1296. https://doi.org/10.1002/aur.2151

Halson, S. L., Peake, J. M., & Sullivan, J. P. (2016). Wearable technology for athletes: Information overload and pseudoscience? *International Journal of Sports Physiology and Performance*, *11*(6), 705–706. https://doi.org/10.1123/IJSPP.2016-0486

Task Force of ESC and NASPE (1996) Heart rate variability: Standards of measurement, physiological interpretation and clinical use. *Circulation*, *93*(5), 1043–1065.

JCS Joint Working Group. (2012). Guidelines for the clinical use of 24 hour ambulatory blood pressure monitoring (ABPM) (JCS 2010). *Circulation Journal*, *76*(2), 508–519. https://doi.org/10.1253/circj.CJ-88-0020

Kasos, K., Zimonyi, S., Gonye, B., Köteles, F., Kasos, E., Kotyuk, E., … Szekely, A. (2019). Obimon: An open-source device enabling group measurement of electrodermal activity. *Psychophysiology*, *56*(8), e13374. https://doi.org/10.1111/psyp.13374

Kleckner, I. R., Wormwood, J. B., Simmons, W. K., Barrett, L. F., & Quigley, K. S. (2015). Methodological recommendations for a heartbeat detection-based measure of interoceptive sensitivity. *Psychophysiology*, *52*(11), 1432–1440. https://doi.org/10.1111/psyp.12503

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs.

*Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL.

Leibold, N. K., & Schruers, K. R. (2018). Assessing panic: Bridging the gap between fundamental mechanisms and daily life experience. *Frontiers in Neuroscience*, *12*, 785. https://doi.org/10.3389/fnins.2018.00785

McCarthy, C., Pradhan, N., Redpath, C., & Adler, A. (2016). Validation of the Empatica E4 wristband. *2016 IEEE EMBS International Student Conference (ISC)*, 1–4. https://doi.org/10.1109/EMBSISC.2016.7508621

Menghini, L., Gianfranchi, E., Cellini, N., Patron, E., Tagliabue, M., & Sarlo, M. (2019). Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiology*, *56*(11). https://doi.org/10.1111/psyp.13441

Michel, V., Mazzola, L., Lemesle, M., & Vercueil, L. (2015). Long-term EEG in adults: Sleep-deprived EEG (SDE), ambulatory EEG (Amb-EEG) and long-term video-EEG recording (LTVER). *Neurophysiologie Clinique = Clinical Neurophysiology*, *45*(1), 47–64. https://doi.org/10.1016/j.neucli.2014.11.004

Mittal, S., Movsowitz, C., & Steinberg, J. S. (2011). Ambulatory external electrocardiographic monitoring: Focus on atrial fibrillation. *Journal of the American College of Cardiology*, *58*(17), 1741–1749. https://doi.org/10.1016/j.jacc.2011.07.026

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Moore, S. T., MacDougall, H. G., & Ondo, W. G. (2008). Ambulatory monitoring of freezing of gait in Parkinson's disease. *Journal of Neuroscience Methods*, *167*(2), 340–348. https://doi.org/10.1016/j.jneumeth.2007.08.023

Mukhopadhyay, S. C. (2015). Wearable sensors for human activity monitoring: A review. *IEEE Sensors Journal*, *15*(3), 1321–1330. https://doi.org/10.1109/JSEN.2014.2370945

Myrtek, M. (2004). *Heart and emotion: Ambulatory monitoring studies in everyday life*. Ashland, OH, US: Hogrefe & Huber Publishers.

Nelson, E. C., Verhagen, T., Vollenbroek-Hutten, M., & Noordzij, M. L. (2019). Is wearable technology becoming part of us? Developing and validating a measurement scale for wearable technology embodiment. *JMIR MHealth and UHealth*, *7*(8), e12771. https://doi.org/10.2196/12771

Parati, G., Stergiou, G. S., Asmar, R., Bilo, G., de Leeuw, P., Imai, Y., … Mancia, G. (2010). European Society of Hypertension practice guidelines for home blood pressure monitoring. *Journal of Human Hypertension*, *24*(12), 779–785. https://doi.org/10.1038/jhh.2010.54

Parati, Gianfranco, Stergiou, G., O'Brien, E., Asmar, R., Beilin, L., Bilo, G., … Zhang, Yuqing. (2014). European Society of Hypertension practice guidelines for ambulatory blood pressure monitoring. *Journal of Hypertension*, *32*(7), 1359–1366. https://doi.org/10.1097/HJH.0000000000000221

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

Peake, J. M., Kerr, G., & Sullivan, J. P. (2018). A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Frontiers in Physiology*, *9*, 743. https://doi.org/10.3389/fphys.2018.00743

Poh, M.-Z., Swenson, N. C., & Picard, R. W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Bio-Medical Engineering*, *57*(5), 1243–1252. https://doi.org/10.1109/TBME.2009.2038487

Quigley, K. S., Barrett, L. F., & Weinstein, S. (2002). Cardiovascular patterns associated with threat and challenge appraisals: A within-subjects analysis. *Psychophysiology*, *39*(3), 292–302. https://doi.org/10.1017/S0048577201393046

Sano, A., Picard, R. W., & Stickgold, R. (2014). Quantitative analysis of wrist electrodermal activity during sleep. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *94*(3), 382–389. https://doi.org/10.1016/j.ijpsycho.2014.09.011

Savard, M.-H., Savard, J., Caplette-Gingras, A., Ivers, H., & Bastien, C. (2013). Relationship between objectively recorded hot flashes and sleep disturbances among breast cancer patients: Investigating hot flash characteristics other than frequency. *Menopause (New York, N.Y.)*, *20*(10), 997–1005. https://doi.org/10.1097/GME.0b013e3182885e31

Scerbo, A. S., Freedman, L. W., Raine, A., Dawson, M. E., & Venables, P. H. (1992). A major effect of recording site on measurement of electrodermal activity. *Psychophysiology*, *29*(2), 241–246.

Schaffer, K., Panneerselvam, N., Loh, K. P., Herrmann, R., Kleckner, I. R., Dunne, R. F., … Fung, C. (2019). Systematic review of randomized controlled trials of exercise interventions using digital activity trackers in patients with cancer. *Journal of the National Comprehensive Cancer Network: JNCCN*, *17*(1), 57–63. https://doi.org/10.6004/jnccn.2018.7082

Schandry, R. (1981). Heart beat perception and emotional experience. *Psychophysiology*, *18*(4), 483–488.

Shcherbina, A., Mattsson, C., Waggott, D., Salisbury, H., Christle, J., Hastie, T., Wheeler, M., & Ashley, E. (2017). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of Personalized Medicine*, *7*(2), 3. https://doi.org/10.3390/jpm7020003

Spagnolli, A., Guardigli, E., Orso, V., Varotto, A., & Gamberini, L. (2014). Measuring user acceptance of wearable symbiotic devices: Validation study across application scenarios. In G. Jacucci, L. Gamberini, J. Freeman, & A. Spagnolli (Eds.), *Symbiotic Interaction* (pp. 87–98). Springer International Publishing.

Stern, R. M., Ray, W. J., & Quigley, K. S. (2000). *Psychophysiological Recording* (2 edition). Oxford England; New York: Oxford University Press.

Straiton, N., Alharbi, M., Bauman, A., Neubeck, L., Gullick, J., Bhindi, R., & Gallagher, R. (2018). The validity and reliability of consumer-grade activity trackers in older, community-dwelling adults: A systematic review. *Maturitas*, *112*, 85–93. https://doi.org/10.1016/j.maturitas.2018.03.016

Strangman, G. E., Ivkovic, V., & Zhang, Q. (2018). Wearable brain imaging with multimodal physiological monitoring. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, *124*(3), 564–572. https://doi.org/10.1152/japplphysiol.00297.2017

van Dooren, M., de Vries, J. J. G. G.-J., & Janssen, J. H. (2012). Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & Behavior*, *106*(2), 298–304. https://doi.org/10.1016/j.physbeh.2012.01.020

van Lier, H. G., Pieterse, M. E., Garde, A., Postel, M. G., de Haan, H. A., Vollenbroek-Hutten, M. M. R., … Noordzij, M. L. (2019). A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the E4 biosensor. *Behavior Research Methods* https://doi.org/10.3758/s13428-019-01263-9

Venables, P. H., & Christie, M. J. (1980). Electrodermal activity. In I. Martin, & P. H. Venables (Eds.), Techniques in psychophysiology (pp. 3-67). New York: John Wiley & Sons.

Weippert, M., Kumar, M., Kreuzfeld, S., Arndt, D., Rieger, A., & Stoll, R. (2010). Comparison of three mobile devices for measuring R-R intervals and heart rate variability: Polar S810i, Suunto t6 and an ambulatory ECG system. *European Journal of Applied Physiology*, *109*(4), 779–786. https://doi.org/10.1007/s00421-010-1415-9

Whitehead, W. E., Drescher, V. M., Heiman, P., & Blackwell, B. (1977). Relation of heart rate control to heartbeat perception. *Biofeedback and Self-Regulation*, *2*(4), 317–392.

Wilhelm, Frank H., & Grossman, P. (2010). Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. *Biological Psychology*, *84*(3), 552–569. https://doi.org/10.1016/j.biopsycho.2010.01.017