

**Title:** Guidelines for Wrist-Worn Consumer Wearables Assessment of Heart Rate in  
Biobehavioral Research

**Authors:** Benjamin W. Nelson, M.S.<sup>a,b</sup>, Carissa A. Low, Ph.D.<sup>c</sup>, Nicholas Jacobson, Ph.D.<sup>d,e</sup>,  
Patricia Areán, Ph.D.<sup>f</sup>, John Torous, M.D.<sup>g</sup>, & Nicholas B. Allen, Ph.D.<sup>a</sup>

**Affiliations:**

- a. Department of Psychology, University of Oregon, Eugene, OR, USA
- b. School of Medicine, University of Washington, Seattle, WA, USA
- c. Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
- d. Geisel School of Medicine, Dartmouth College, Hanover, NH, USA
- e. Center for Technology and Behavioral Health, Dartmouth College, Hanover, NH, USA
- f. Department of Psychiatry, University of Washington, Seattle, WA, USA
- g. Beth Israel Deaconess Medical Center, Department of Psychiatry, Harvard Medical  
School, Boston, MA, USA

**Corresponding Author:**

Benjamin W. Nelson  
Department of Psychology  
1227 University of Oregon  
Eugene, OR 97403 USA  
Email: [bwn@uoregon.edu](mailto:bwn@uoregon.edu)

## **Abstract**

Researchers have increasingly begun to use consumer wearables or wrist-worn smartwatches and fitness monitors for measurement of cardiovascular psychophysiological processes related to mental and physical health outcomes. These devices have strong appeal because they allow for continuous, scalable, unobtrusive, and ecologically valid data collection of cardiac activity in “big data” studies. However, replicability and reproducibility may be hampered moving forward due to the lack of standardization of data collection and processing procedures and inconsistent reporting of technological factors (e.g., device type, firmware versions, sampling rate), biobehavioral variables (e.g., body mass index, wrist dominance and circumference), and participant demographic characteristics, such as skin tone, that may influence heart rate measurement. These limitations introduce unnecessary noise into measurement, which can cloud interpretation and generalizability of findings. This paper provides a brief overview of research using commercial wearable devices to measure heart rate, reviews literature on device accuracy, and outlines the challenges that non-standardized reporting pose for the field. We also discuss study design, technological, biobehavioral, and demographic factors that can impact the accuracy of the passive sensing of heart rate measurements, and provide guidelines and corresponding handouts for future study data collection and design, data cleaning and processing, analysis, and reporting that may help ameliorate some of these barriers and inconsistencies in the literature.

## Heart Rate Research in Psychology and Medicine

Heart rate (HR) has long been used as a clinical indicator of overall cardiac health. HR is predominantly influenced by the coordination of the sympathetic and parasympathetic branches of the autonomic nervous system, which can be modified by biopsychosocial factors including physical and relational stress, diet, physical fitness, medication use, and substance use<sup>1</sup> thus offering a direct and quantifiable connection to the stress diathesis model of health. In this model, cardiovascular dysfunction has been proposed to be a putative mechanism associated not only with morbidity and mortality, but also with a range of psychiatric disorders.

In terms of physical health, meta-analyses have shown that resting HR above 80 beats per minute (bpm) is associated with a 33% increased risk for cardiovascular mortality and a 45% higher risk for all-cause mortality<sup>2,3</sup>. In terms of HR dynamics, delayed HR recovery after a stressor (i.e., the return to resting or baseline HR) has also been shown to be associated with mortality<sup>4</sup>, while the literature is more mixed on the association between HR reactivity (i.e., the increase in HR from resting or baseline due to psychobiological stressors) and health outcomes<sup>5</sup>. Furthermore, dysfunction of the autonomic nervous system is also associated with a range of psychiatric disorders<sup>6,7</sup> including anxiety disorders<sup>8</sup>, depressive disorders<sup>9</sup>, posttraumatic stress disorder<sup>10</sup>, and schizophrenia<sup>11</sup>. In a large prospective Swedish study of over 1 million male participants, elevated resting HR during late adolescence was found to be associated with an increased risk of obsessive-compulsive disorder, anxiety diagnosis, and schizophrenia, while lower resting HR was associated with substance use disorders and violent criminality<sup>12</sup>.

In addition to these studies, there is strong evidence that psychiatric diagnoses are associated with increased risk for physical morbidity and mortality. For example, those with psychopathology lose a median of 10 years of life<sup>13</sup>, with the most common cause of death due

to cardiovascular diseases <sup>6</sup>. Cardiac activity therefore may be a biological factor that serves as a mechanism linking mood and anxiety disorders with cardiovascular and metabolic risk and mortality <sup>14,15</sup>. In other words, HR may be a transdiagnostic biomarker with clinical utility for psychology, psychiatry, and medicine to improve nosology and increase identification of those at risk for the onset of psychiatric and physical health disorders. Until recently, the quantification and measurement of HR has been relegated to medical and research settings, which has largely precluded the observation of HR during patient and participant daily life, thereby limiting the generalizability of findings to real-world living conditions.

### **Current Limitations in Cardiovascular Research**

The foundation of HR research has, until recently, been largely constrained due to 1) lack of ecological validity (i.e., the degree to which a laboratory or medical environment represents actual real-world conditions), 2) inability to collect temporally detailed and longitudinal HR measurements both within and across time (e.g., hours, days, weeks, months), and 3) prohibitive financial cost of devices for everyday use.

In terms of ecological validity, HR recordings have historically been collected in discrete laboratory or medical environments that do not represent daily life and as such, can be anxiety-inducing as medical phobia is common and novel environments may increase HR, which consequently could lead to overestimation of resting HR <sup>16</sup>, which has been found in greater adrenergic activity in clinical settings, which is a known phenomenon called “white-coat hypertension or syndrome” <sup>17,18</sup>. In other words, research has been historically limited to collecting data from medical patients and research participants in novel settings that might not generalize to their everyday lives, which creates the potential for inaccurate baselines and therefore inaccurate measurement of HR dynamics (i.e., reactivity/recovery). The Law of Initial

Values gives further insight into this issue by highlighting that an initial value has a large impact on the strength and direction of a response, such that a higher pre-stimulus level will lead to a smaller post-stimulus response or, in other words, a smaller response when there is a greater initial value<sup>19-21</sup>. This is similar to a ceiling effect limiting the magnitude of a reactivity response and this can artificially constrain the range of measurements (e.g., reactivity) within a specific construct, such as HR.

Second, collection of HR data in these environments lacks temporal resolution within and across days as HR is usually collected during a short recording within a single day, which has precluded research from comprehensively investigating the transition from acute to chronic stress<sup>22</sup>. Greater temporally detailed longitudinal HR data collection may reveal important within day, daily, weekly, monthly, seasonal, and yearly fluctuations in HR that are currently not captured and may contain important diagnostic and predictive information. Lastly, the accurate collection of HR has historically been financially restrictive relegating these measurements to medical and scientific contexts, and precluding the large-scale adoption and democratization of these technologies by the general public.

### **Wearable Technological Advancements and The Scalability of Cardiovascular Research**

Currently, wrist-worn wearables provide an opportunity to examine HR in real-world environments, across longer periods of time, and with high resolution at a low cost, by using an optical sensor (photoplethysmography; PPG), that allows these devices to collect volumetric changes in blood perfusion (i.e., pulse rate) that can serve as a surrogate for HR. Wearable PPG sensors can use different colors of light, such as green or red, to index these changes in blood perfusion with each having different costs and benefits. For example, green LED light is a shorter wave-length and has been shown to be convenient, because it has good signal-to-noise

ratio and is resistant to motion artifacts, but is limited by the amount of light that can pass through tissue, especially darker skin tones. In contrast, red/infrared LED, which are more commonly used in medical settings, is not absorbed as well by the skin allowing the light to transmit deeper and allowing for detection of multiple biological measures, but it has a lower signal-to-noise ratio and is more susceptible to motion artifacts<sup>23</sup>. Overall, PPG devices allow for *in vivo* examinations of the roles that resting HR and HR dynamics (e.g., reactivity and recovery) play in medical and psychiatric disorders in real-world settings. It is important to note that while HR and pulse rate are often used interchangeably, they are two distinct physiological signals with HR representing the heart contraction via electrical impulses, while pulse rate represents the rate of change in blood pressure due to the ventricular ejection of blood. The latter of these two is the predominant method for data collected with wrist-worn wearable devices, although some more recent wrist-worn wearable devices, such as the Apple Watch Series 4 and 5, have allowed for the discrete, rather than continuous, collection of HR via electrical impulses on par with a single-lead electrode. Currently, this technology is limited, as users have to sit still with their arm resting on their lap or table and put a finger from the hand opposite to the watch to close the circuit for 30 seconds, therefore precluding the ability for temporally detailed longitudinal ECG collection at this time. For this reason, addressing accuracy of wearable ECG sensors is outside the scope of the current review and will not be discussed further. Furthermore, PPG can be used to index heart rate variability (HRV), blood pressure, oxygen saturation, and cardiac output<sup>24</sup>, but as of yet have not been widely rolled out in consumer-use devices and therefore data collection other than HR will not be discussed further.

## **Increased Popularity of Wrist-Worn Wearable Devices**

Researchers are increasingly using consumer wearables or wrist-worn smartwatches and fitness monitors for the digital measurement of cardiovascular health and psychophysiological stress as these devices allow for continuous, scalable, and unobtrusive, “big data” collection of overall cardiac activity in real-world conditions with large samples<sup>25-30</sup>. Between 2008 and 2018 there was a 580% increase in total articles published that contained “wearable” and “heart rate”, indicating that this is a burgeoning area of research that may benefit from standardization of procedures to increase generalizability and clinical utility of research findings.

Recent information on the reproducibility and replicability crisis has swept the sciences and as such these issues are very important for digital health and mHealth research, which includes wrist-worn HR accuracy and any studies that utilize wearable HR data, as this new field will allow for large data samples<sup>31</sup> with various sensors that will increase the ease for p-hacking. Recent research has proposed using wearables in psychological research<sup>32</sup>, treatment settings<sup>33</sup>, while other research has already begun to use mHealth data to show important global real-world HR norms<sup>16</sup> and global differences in HRV by age, gender, and time of day<sup>26</sup>, and allowing for surveillance of influenza symptoms<sup>34</sup>, yet questions remain as to the accuracy and validity of data collected from wearables<sup>35,36</sup>.

## **How Accurate is HR Detection with Wrist-Worn Wearable Devices?**

Current research into the HR accuracy of wrist-worn wearable devices attempt to compare the measurement between wearable PPG and a reference method, such as an electrocardiogram (ECG), in order to determine whether or not measurements between the devices are outside of clinically important limits of agreement, so that devices can supplement, eventually replace, or be used interchangeably<sup>37,38</sup>. While some studies have attempted to address

wearable HR accuracy using the gold-standard reference method of ECG, which directly measures HR via electrical impulses from the heart and is used for clinical accuracy,<sup>38-44</sup> other studies have used suboptimal reference methods, such as the chest straps<sup>45-47</sup> or pulse oximeters<sup>48</sup>, which themselves have a degree of error (varying between 0 – 40 bpm when compared to ECGs<sup>44</sup>). This has the potential to introduce a degree of additional error into wrist-worn wearable HR accuracy research as suboptimal referencing methods can introduce unnecessary measurement error and potentially undermine findings. Despite this caveat, current research into wearable HR accuracy is promising as is discussed below. Furthermore, some wearable devices, such as the Apple Watch 4 and 5, have gained FDA 510(k) class II clearance for the ECG feature and ability to detect arrhythmias<sup>49</sup>, indicating that FDA clearance and conformance with IEC-60601(-2-47) guidelines may provide a path towards standardization of feature sets and firmware for consumer wearables in the future, which raises the important question of the interface between consumer wearables and medical devices. Currently, in most instances, consumer wearables are not medical devices, but this is an issue that is evolving and in the future may influence legislation and standardization. Below we mainly focus on HR accuracy of Apple Watch and Fitbit as these are two of the most popular wearable companies in the United States and have been more commonly examined in empirical research to date, but it should be noted that other consumer wearable brands, such as Garmin<sup>45,46,50</sup>, Samsung<sup>42,43,48,51</sup>, and Microsoft<sup>42,47</sup> have also had studies examining their HR accuracy.

Current research verifying wearables to the gold-standard ECG, which uses electrodes to directly measure cardiac muscular contractions from electrical activity of the heart, have shown that on average wearables slightly and negligibly underestimate (-0.9 to -7.2 bpm) absolute HR<sup>40,41,44,52</sup> with the Apple Watch having slightly greater accuracy than Fitbit devices (Apple Watch:



-1.3 bpm, Fitbit: -9.3 bpm<sup>42-44</sup>, as well as lower overall error with the Apple Watch models ranging from 1.20 - 6.70% error and Fitbit models ranging from 2.38 - 21.36% error<sup>39,42,45</sup>, lower mean difference standard deviation<sup>43</sup>, and higher agreement with ECG than Fitbit devices<sup>39,44</sup>. Furthermore, recent research has indicated that consumer wearables, such as the Apple Watch and Fitbit are more accurate than research-grade wearables, which are much more expensive yet provide the benefit of raw data<sup>50</sup>. These findings have led some researchers to conclude that wearables detect HR with acceptable accuracy in both laboratory and real-world settings for research, but importantly not medical settings<sup>38,42</sup> as more research needs to be done. For example, researchers compared the accuracy of ambulatory ECG, Apple Watch 3, and Fitbit Charge 2 data within an individual across 24 hours and found that while devices slightly underestimated HR when values were aggregated within and across different daily activities, which indicated high HR accuracy, any individual HR observation could deviate from ECG by significantly large margins that would be likely be problematic in medical settings<sup>38</sup>. This finding indicated that while overall, summary statistics after outlier detection and data cleaning may be very accurate for research purposes, any single observation in real-time may have a large degree of error, which could be significant for moment-to-moment observations in medical settings. These preliminary findings suggest that researchers can utilize wearable HR to accurately assess HR, especially resting HR, when using aggregated data across activities and removing outliers. The next steps for consumer wearable HR device accuracy research will require meta-analyses in order to summarize the current state of consumer wearable HR accuracy. Despite the promise of wearable HR accuracy, replicability and reproducibility in the future utilizing wearable HR as a transdiagnostic marker of psychiatric and physical health will likely be limited due to lack of standardization of data collection, study design, and data

processing as well as inconsistencies in the reporting of technological, biobehavioral, and demographic characteristics that introduce unnecessary noise into the literature and cloud interpretation and generalizability of findings.

### **Disadvantages to Non-Standard Reporting**

The lack of “metadata standards” or guidelines and standards for describing, reporting, and creating meaning from collected data in order to increase transparency, accountability, and interpretability of data in mHealth is likely to create a number of issues related to noise/signal ratio, contradictory results, lack of generalizability, and false positive and negative findings, which have historically undermined replicability (i.e., the ability to replicate a study with a new dataset) and reproducibility (i.e., the ability to replicate findings from the same dataset).

Therefore, many researchers have begun to push for increased focus on reproducible and open practices in science<sup>31,53–55</sup>. Below we outline the data collection, cleaning, and protocol design issues as well as technological, biobehavioral, and demographic factors that may influence reproducibility and replicability before turning to guidelines that the field can use in order to increase scientific rigor and generalizability of findings.

### **Study Protocol Design and Data Collection: Addressing Design, Biobehavioral, and Demographic Issues That Can Influence Reproducibility and Replicability**

There are many researcher degrees of freedom when it comes to decisions related to study protocol design and data collection, cleaning, and processing that are important to note as standardizing this process may help ameliorate some barriers and inconsistencies in the literature.

#### *Protocol Design.*

*Summary Metric Calculation.* Hardware differences between devices determines HR measurement second to second, which has large implications on overall interpretability of the data, such that sampling rate can introduces researcher degrees of freedom for calculating beats per minute (bpm). Unfortunately, there is currently a lack of transparency from manufacturers on underlying algorithms and raw data outputs are often times not provided, which prevents researchers from dealing with this issue and creating a standardized protocol for transforming raw data to HR. For example, Fitbit logs HR data at 1 second intervals during exercise tracking and 5 second intervals at other times, while the Apple Watch samples HR data every 10 minutes, except during workouts, which are continuous. Relatedly, some devices, such as the Apple Watch can take multiple HR measurements within a minute allowing researchers to use various calculations, such as the mean or median within this minute to calculate bpm, which may alter wearable data values. Researchers should report how bpm or other summary metrics are calculated (e.g., mean or median). Therefore, summary metric calculation should also be clearly reported in publications.

*Data Cleaning: Outlier Detection and Dealing With Missing Data.* In addition, the methods researchers select to clean and process raw data prior to analyses can also affect results and conclusions and should be reported in detail. These choices include how missing data values are handled (listwise deletion, pairwise deletion, multiple imputation) and how outliers or biologically implausible values are handled (e.g., some devices record HR values of zero when devices are not worn).

Data cleaning can be an important step in data pre-processing in processing ambulatory HR assessments. There is a strong need to customize the methods of data cleaning to the device being used and the models being adopted, rather than adopting a universal pre-processing

pipeline due to device-specific differences and based on the study research questions. For example, outliers in linear models may be strongly impact linear estimators, and some studies may therefore adopt a modeling strategy wherein values are winsorized (e.g., extreme values can be truncated at the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles <sup>56</sup>). However, if a study were interested in HR arrhythmias using wearable assessments (such as atrial fibrillation), then these extreme values may be of vital interest and should not be removed or altered <sup>27</sup>. Moreover, for some, but not all devices will report missing HR data with zero values and these values should be removed and filled in with missing data rather than winsorized (which would otherwise substantially bias parameter estimates unless the degree of missing HR points was trivial). Other strategies (e.g. spectral frequency filters), will depend upon the type of the sensor being utilized in the wearable sensor <sup>57</sup>. Rather than proposing universal standards, we rather recommend that researchers thoughtfully approach data cleaning, altering their decisions to both their own research questions and to the devices being used, and reporting their utilized methodology.

The practice of completely removing persons from studies due to proportions of missing data (i.e. listwise deletion) is known to bias estimation and adversely impact the standard error of point estimates across modeling strategies <sup>58</sup>. More specifically, simulation studies with intensive longitudinal data suggest that multiple imputation and full-information maximum likelihood reliably produce better parameter estimates than listwise deletion <sup>59</sup>. Moreover, dynamic models of HR captured in daily life can be effectively estimated even with very large proportions of missing data (e.g. greater than 70% missingness for each subject) using full-information maximum likelihood estimation <sup>60</sup>. Based on this, we recommend that authors aim to be as inclusive as possible, using appropriate model-based strategies for dealing with missing data.

*Defining Non-Wear Time: Device Charging and Non-Adherence.* For long-term field studies, information about participant adherence and wear-time should be considered and reported since these issues can affect data quality. The definition of wear time and any minimum wear-time thresholds for data inclusion (e.g., days with wear time less than 1000 minutes excluded; Radin et al., 2020; valid day of wear time defined as at least 600 1-minute epochs of nonzero HR values within a calendar day) <sup>61</sup>. One issue reported by multiple users on the online Fitbit Community Forum is occasional logging of HR values (often higher and more variable than typical, e.g., 100-150 bpm) when the devices are not being worn, suggesting that supplemental non-wear time diaries may be required for some devices or protocols.

*Standardizing Wrist Data Feature Collection.* There are a number of study design features related to the wrist that have been shown to have effects on HR accuracy. Larger wrist circumference has been shown to be associated with reduced HR accuracy<sup>42</sup>, although these outcomes have not been found in all studies <sup>44,46</sup>. In addition to wrist circumference, there are a number of other factors including wrist placement, tightness of device, dominant vs non-dominant hand use, and degree of wrist movement that have consistently been shown to influence the accuracy of HR measurement <sup>39</sup>, which may be due to the fact that greater wearable movement is associated with decreased PPG accuracy as discussed below <sup>62</sup>. These are likely some of the most important factors influencing HR accuracy. As such, wrist placement, tightness of device, dominant vs non-dominant hand use, and whether wearable devices are naturalistically placed by participants as opposed to explicit instructions by experimenters should be reported.

*Open Science and Data Transparency.* Lastly, from a study design perspective, mHealth research will be particularly susceptible to a number of research practices, which have been shown to degrade the quality, consistency, generalizability, and therefore replicability and

reproducibility of findings<sup>31</sup> as data sensors in this research field allow for the vast collection of big data samples, which will be vulnerable to some of the “four horsemen of the reproducibility/replicability apocalypse,” including 1) publication bias<sup>63</sup>, 2) P-hacking, 3) hypothesizing after the results are known or HARKing<sup>54</sup>, and 4) low power<sup>64</sup>, the last of which likely will not be an issue for mHealth research as large samples sizes will be relatively easy to collect even in samples with low base rates in the population. Therefore, we urge future wearable HR studies and digital passive sensing mHealth studies in general to take up free open science practices (<http://osf.io/>), such as study preregistration, open code, and open data and where open data cannot be released due to clinical population or privacy issues, releasing simulated datasets that preserve dataset characteristics where applicable<sup>65</sup>. On the editorial and reviewer side, we join other researchers in urging for editors and reviewers to allow for null, inconclusive, and contradictory data results without putting undue pressure on researchers to come up with coherent narratives that might not fit the data<sup>66</sup>.

### *Biobehavioral and Demographic Data*

Individual differences and variability in both biobehavioral and demographic factors have the potential to influence the collection of HR during research studies and should therefore be strongly considered in descriptive reporting as well as statistical modeling in studies using HR measurement collected from wearable devices.

*BMI and Biological Sex.* Higher body mass index, which likely covaries with larger wrist circumference as described above, has been shown to be associated with reduced HR accuracy<sup>42,62</sup>, although again these outcomes have not been found in all studies<sup>44,46</sup>. Nevertheless this variable should be collected, reported, and potentially controlled for in research. In addition, research has not always reported participant biological sex, but research has found higher device

error in males compared to females<sup>42</sup>, indicating that participant biological sex should be collected and possibly controlled.

*Skin Tone and Hair Follicle Density.* Evidence is starting to accumulate on the importance of skin tone when it comes to accuracy and generalizability of wrist-worn wearable studies with darker skin tone (and tattoos) being found to be associated with reduced wrist-worn HR accuracy as they absorb more green light<sup>67</sup>, yet this is not found in all studies<sup>43,50</sup>. The importance of this point cannot be overstated as historically, science and medicine have paid less attention to recruiting minority groups<sup>68</sup>, which has resulted in blunted health gains in these populations making it of utmost importance to identify how HR accuracy varies by race and ethnicity factors in order to ensure that findings generalize to these populations as future research might have direct implication for patient healthcare. Currently, the vast majority of wrist-worn wearable studies incorporating HR measurements do not report or control for participant race and ethnicity, which undermines the ability to look at potentially important group differences in health outcomes and therefore decreases the ability of results to generalize to large segments of the population. These demographic characteristics should be collected, reported, and controlled for in all wearable HR studies moving forward and researchers should collect data related to participant skin tone (e.g., Fitzpatrick Skin Type). Furthermore, wearables that use red light will likely be more effective on darker skin tones as it is not absorbed by melanin, which may allow for more accurate heart rate measurements<sup>23</sup>. In addition to skin tone, some research indicates that hair follicle density and sweat can influence device measurement<sup>69</sup>.

*Motion Artifacts due to Physical Activity.* Importantly, two related factors identified as influencing wearable HR accuracy are motion artifacts<sup>23</sup> and level of physical activity<sup>46</sup>. Specifically, research has shown that at rest, wearables can perform similarly to an ECG, with

some research showing that absolute error during activity is 30% higher than rest<sup>50</sup> and a substantial amount of research showing that wearable devices are more accurate during rest and low intensity exercise when compared to higher intensity exercise, when wearable devices can begin to deviate more from ECG recordings<sup>38,39,44,46,70-73</sup>, although this has not been found in all studies<sup>42,45,47,51</sup>, with some studies finding that accuracy of HR measurement was comparable across resting baseline or vigorous activity<sup>45</sup>, while a second study found that the accuracy of HR measurement was *highest* during running<sup>51</sup>, and a third found that measurement during walking, running, and cycling was more accurate for some devices than during sitting<sup>42</sup>. Other research found that across 4 consumer wearables and 2 research grade devices are all were “reasonably accurate” during rest or during prolonged heart rate that was elevated and that differences in accuracy tended to arise during changes in activity<sup>50</sup>. Therefore, it is possible that activity intensity may be less important to device accuracy than the degree of erratic wrist movements and corresponding position of the wearable device during an activity<sup>74</sup>, which cause motion artifacts and may or may not co-occur with more vigorous physical activity. Furthermore, research has described the “cross-over effect” when repetitive motion and activity causes underlying algorithms to mistake the movement cadence for cardiac activity<sup>50,75</sup>. As mentioned previously one strength of PPG green light is that it is less susceptible to motion artifacts, while PPG red light is more susceptible to motion artifacts indicating that wearables that use green PPG or a combination of green and red PPG are likely to perform better during movement and physical activity<sup>23</sup>. Future studies may want to control for wearable actigraphy data in order to remove HR accuracy variance that is influenced by wrist movement.



### *Technological Factors That Can Influence Reproducibility and Replicability*

In addition to the high heterogeneity in study protocol design as described above, various technological factors have the potential to influence the collection of HR during research studies and should therefore be strongly considered in descriptive as well as statistical modeling when using wearable devices. Specifically, due to the variety of wearable devices within and across technology companies, versions of devices, and the fast pace of technological advancement (especially in consumer products) there are a number of technological factors that should be reported and potentially controlled for in mHealth studies. First, as described above, there are currently various wearable devices and versions on the market that have the potential to use different hardware, which may introduce differences in HR recordings. One such hardware difference between devices are sensor type. For example, the Apple Watch uses both PPG green light-emitting diode (LED) lights and infrared LED light, while the Fitbit uses only PPG green LED.

In addition to hardware differences, researchers should also consider software differences between devices. Wearable manufacturers use proprietary algorithms to translate PPG signals to HR measurements, which may influence accuracy<sup>23</sup>. These algorithms, while often simplistic, may be altered with firmware updates, yet most studies fail to report firmware information. Not controlling for firmware version may lead to poorer reproducibility and replicability as two studies investigating the same device with two or more different firmware versions might actually come to different conclusions even if all other variables are held constant.

Due to the high heterogeneity in device type, version, hardware, software, and sampling rate, these should be consistently collected and reported in descriptive tables for both cross-sectional and longitudinal studies. In addition, if a study uses multiple types of devices or a

single device in a longitudinal study that covers a period of time when there are firmware updates, then statistical models should control for device type and firmware version.

## **Guidelines**

Here we join other researchers in the call for the establishment of “metadata standards” or guidelines for collecting, processing, describing, reporting, and creating meaning of collected HR and other biological data in order to increase transparency, accountability, and interpretability of data<sup>28,76,77</sup>. With this in mind we have created two checklists. The first checklist addresses standards for reporting study characteristics, including how the study was designed, technological hardware and software used; participant characteristics; and data cleaning, analysis, reporting, and transparency<sup>1</sup>. The second checklist provides potential covariates in regards to technological factors and both biobehavioral and demographic factors that have also been shown to influence accuracy of HR measurements and dynamics. We propose that future research utilizing wearable HR should report in methods sections that study design protocol, biobehavioral, participant demographic, and technological variables have been reported in accordance with these standardized guidelines and include these checklists in supplemental materials. Utilizing these guidelines and referencing them in methods sections has the potential to decrease non-standard reporting, increase reproducibility and replicability, while also leading to greater generalizability of results, which may increase the clinical utility of findings in this field.

---

<sup>1</sup> Note that we suggest that researchers report the reliability of their metric and justify why this reliability threshold is sufficient for the given study. Note that we make this recommendation purposefully and do not adopt a single global threshold for all of research. We do this because research is heavily context dependent. For example, a very high standard might be needed when heart rate is used to classify heart disease. However, in contrast to this situation, when heart rate is only one of many potential avenues of information it might be sufficient that heart rate signal is reliably better than complete noise. For instance, adopting a merely better-than-noise standard in a multi-component machine learning model this could still improve performance and may be preferable when researchers are focused on scalable low-cost devices. In the second scenario, imposing an arbitrary standard might make a model less accurate because it might otherwise require the researchers to drop heart rate signal entirely.

## **Conclusion**

HR has long been used as a clinical indicator of overall cardiac health that has been conceptualized as a transdiagnostic biomarker with clinical utility in psychology, psychiatry, and medicine to improve nosology and increase identification of those at risk for the onset and relapse of psychiatric and physical health conditions. Cardiovascular dysfunction has been proposed to be a putative mechanism associated not only with morbidity and mortality, but also with a range of psychiatric disorders. Until recently, real-world continuous collection of HR in medical patient and research participant lives has been limited by technological and financial factors. The recent introduction of consumer wearables or wrist-worn smartwatches and fitness monitors for the digital measurement of cardiovascular health and psychophysiological stress has allowed for the continuous, scalable, and unobtrusive, “big data” collection of overall cardiac activity in real-world conditions with large samples that allow for the collection of passive and large scale collection of ecologically valid data. This has the potential to improve the study of cardiovascular health in medicine and psychophysiological stress in psychiatry and psychology.

The current literature indicates that these devices can provide acceptable accuracy for the measurement of HR for research settings, especially when scalability and ecological validity of measurement settings are critical considerations that may outweigh the need to offer gold standard assessment. It is important to note that PPG technology can be used to index blood pressure, oxygen saturation, and cardiac output, which will likely be slowly rolled out into wearable technology moving forward. In addition PPG can index HRV, which is derived from beat to beat intervals calculations in the time-domain, such as root mean squared of successive differences or the standard deviation of NN intervals, the latter of which is used in the Apple Watch, but only with sporadic recordings throughout the day. Currently, many commercial

wearables don't provide beat to beat interval data to users, although some researchers have worked with industry partners to derive measurements of HRV <sup>26</sup>. The guidelines provided above will be necessary to apply to research when studying PPG data at more fine grained detail than HR.

Given the variations in both hardware and software in these devices, variation in study design protocol, as well the way in which individual differences across both biobehavioral and demographic characteristics may affect the accuracy of measurement, the field requires standards of reporting that will at the very least allow for the characterization of these factors when comparing studies. In this paper we have proposed a pair of reporting checklists that indicate the details that should be included in any paper using this form of HR measurement.

Overall, while consumer wearable devices that collect HR likely are not yet advanced enough to be replacing medical grade ECG anytime soon, these devices can be used to supplement medical grade devices with medical patients and continuously collect HR data on research subjects when risk for acute cardiac events is not of immediate concern. This has the potential to advance research into cardiovascular health and psychophysiological stress in order to better understand how HR influences overall physical and psychological health.

Table 1. Summary of Most Popular Consumer Wearable Devices

Device*	Sensor Type			FDA Status	Sampling Rate (i.e., how often it samples)	Sampling Frequency	Cost	Market Size/Share 2019 Q4 <sup>78</sup>
	Green LED	Red/IF LED	ECG					
Apple <sup>79</sup>	X	X	X	510(k) class II clearance <sup>a</sup>	Variable <ul style="list-style-type: none"> <li>Rest- every 10 minutes</li> <li>Exercise- continuous</li> </ul>	100s x per second	\$199 (v3) to \$399 (v5)	36.5%
Fitbit <sup>80</sup>				None	Variable <ul style="list-style-type: none"> <li>Rest- 5 second intervals</li> <li>Exercise- 1 second intervals</li> </ul>	Unknown	\$149 (Charge 4) to \$199 (Versa 2)	5.0%
Garmin <sup>81</sup>	X			None	Variable <ul style="list-style-type: none"> <li>High Frequency</li> <li>Low Frequency</li> </ul>	Variable	\$129 (vivosmart 4)	*
Samsung	X			None	Manual	Unknown	Galaxy Fit (\$99)	8.8%
Xiaomi	X			None	Continuous	Unknown	Mi Band 4 (\$39.99)	10.8%
Huawei <sup>82</sup>	X			None	Default is set to manual. Can turn on continuous, which measures every 10 minutes or every 6-10 seconds during high-intensity workouts	Unknown	Huawei Band 4 (\$42.31)	7.8%
<p>*Note: Data is presented on most recent wearable devices for each manufacturer. Garmin was not included in the top 5, so was listed under Other</p> <p><sup>a</sup>Note: This only applies to the ECG sensor and high heart rate notifications.</p>								

mHealth Wearable Heart Rate Metadata Checklist 1:  
Descriptive Reporting of Sample

Describe

Study Design  
Protocol

- Naturalistic or Laboratory
- Group  
(Psychiatric group diagnostic or symptom selection criteria)
- Recruitment Source
- Inclusion/Exclusion Criteria  
(Presence or absence of conditions, age range)

Technological  
Factors

- Device Manufacturer
- Device Type
- Device Version
- Firmware Version
- Hardware (sensor type)
- Sampling Rate
- Device reliability and justification for why this level of reliability is adequate for the study design

Participant  
Characteristics

- Age
- Race
- Ethnicity
- Biological Sex
- Gender
- Skin tone (e.g., Fitzpatrick Skin Type)
- Body Mass Index
- Wrist Circumference
- Wrist Placement (e.g., dominant or non-dominant)
- Medical Condition
- Cardioactive Medication Use

Data Cleaning,  
Handling, and  
Analysis

- Summary bpm Metric Calculation (e.g., mean, median)  
For multiple samples per minute, how was final bpm calculated
- Definition of Non-Wear Time and Reason for data loss (battery life, device failure, participant attrition)
- Dealing with missing data
  - Listwise Deletion (not recommended)

- Pairwise Deletion (not recommended)
- Model-Based (e.g. full-information maximum likelihood, multiple imputation)
- Outlier identification and correction
- Wrist Circumference
- Wrist Placement
  - Dominant or non-dominant
  - Tight vs loose
  - Naturalistic use by participants vs explicit instructions by experimenters

#### Data Reporting

- Reporting Descriptives (e.g., mean, standard deviation, range for overall sample and by group)

#### Data Transparency

- Preregistration
 

Large passive sensing datasets will be ripe for p-hacking. Pre-register analyses prior to viewing data collected or try to draw predictions from out of sample predictions.
- Open Code and Data
 

Provide access to code and data if applicable

mHealth Wearable Heart Rate Metadata Checklist 2:  
Potential Covariates

Control

Technological  
Factors

- Device Manufacturer
- Device Type
- Device Version
- Firmware Version
- Hardware (sensor type)
- Sampling Rate
- Device reliability and justification for why this level of reliability is adequate for the study design

Participant  
Characteristics

- Age
- Race
- Ethnicity
- Biological Sex
- Gender
- Skin tone (e.g., Fitzpatrick Skin Type)  
(e.g., Fitzpatrick Scale)
- Body Mass Index
- Wrist Circumference
- Wrist Placement
  - Dominant or non-dominant
  - Tight vs loose
  - Naturalistic use by participants vs explicit instructions by experimenters
- Medical Condition
- Cardioactive Medication Use



## References

1. Cacioppo JT, Tassinary LG, Berntson GG, eds. *Handbook of Psychophysiology*. 4th ed. Cambridge University Press; 2017. doi:10.1017/9781107415782
2. Khan H, Kunutsor S, Kalogeropoulos AP, et al. Resting Heart Rate and Risk of Incident Heart Failure: Three Prospective Cohort Studies and a Systematic Meta-Analysis. *Journal of the American Heart Association*. 2015;4(1). doi:10.1161/JAHA.114.001364
3. Zhang D, Shen X, Qi X. Resting heart rate and all-cause and cardiovascular mortality in the general population: a meta-analysis. *CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2015;cmaj.150535. doi:10.1503/cmaj.150535
4. Qiu S, Cai X, Sun Z, et al. Heart Rate Recovery and Risk of Cardiovascular Events and All-Cause Mortality: A Meta-Analysis of Prospective Cohort Studies. *Journal of the American Heart Association*. 2017;6(5). doi:10.1161/JAHA.117.005505
5. Low CA, Salomon K, Matthews KA. Chronic life stress, cardiovascular reactivity, and subclinical cardiovascular disease in adolescents. *Psychosom Med*. 2009;71(9):927-931. doi:10.1097/PSY.0b013e3181ba18ed
6. Alvares GA, Quintana DS, Hickie IB, Guastella AmJ. Autonomic nervous system dysfunction in psychiatric disorders and the impact of psychotropic medications: a systematic review and meta-analysis. *Journal of psychiatry & neuroscience*. 2015;40(6):1–16. doi:10.1503/jpn.140217
7. Kandola A, Ashdown-Franks G, Stubbs B, Osborn DPJ, Hayes JF. The association between cardiorespiratory fitness and the incidence of common mental health disorders: A systematic review and meta-analysis. *Journal of Affective Disorders*. 2019;257:748-757. doi:10.1016/j.jad.2019.07.088
8. Kemp AH, Brunoni AR, Santos IS, et al. Effects of depression, anxiety, comorbidity, and antidepressants on resting-state heart rate and its variability: An ELSA-Brasil cohort baseline study. *American Journal of Psychiatry*. 2014;171(12):1328–1334. doi:10.1176/appi.ajp.2014.13121605
9. Kemp AH, Quintana DS, Quinn CR, Hopkinson P, Harris AWF. Major depressive disorder with melancholia displays robust alterations in resting state heart rate and its variability: Implications for future morbidity and mortality. *Frontiers in Psychology*. 2014;5(NOV):1–9. doi:10.3389/fpsyg.2014.01387
10. Paulus EJ, Argo TR, Egge JA. The Impact of Posttraumatic Stress Disorder on Blood Pressure and Heart Rate in a Veteran Population. *Journal of Traumatic Stress*. 2013;26(1):169–172. doi:10.1002/jts.21785

11. Clamor A, Hartmann MM, Köther U, Otte C, Moritz S, Lincoln TM. Altered autonomic arousal in psychosis: An analysis of vulnerability and specificity. *Schizophrenia Research*. 2014;154(1-3):73–78. doi:10.1016/j.schres.2014.02.006
12. Latvala A, Kuja-Halkola R, Rück C, et al. Association of Resting Heart Rate and Blood Pressure in Late Adolescence With Subsequent Mental Disorders. *JAMA Psychiatry*. 2016;41(2):89–104. doi:10.1001/jamapsychiatry.2016.2717
13. Walker ER, McGee RE, Druss BG. Mortality in Mental Disorders and Global Disease Burden Implications. *JAMA Psychiatry*. 2015;72(4):334–341. doi:10.1001/jamapsychiatry.2014.2502
14. Kiecolt-Glaser JK, Wilson SJ. Psychiatric Disorders, Morbidity, and Mortality: Tracing Mechanistic Pathways to Accelerated Aging. *Psychosomatic medicine*. 2016;78:772–775. doi:10.1097/PSY.0000000000000373
15. Nemeroff CB, Goldschmidt-Clermont PJ. Heartache and heartbreak—the link between depression and cardiovascular disease. *Nat Rev Cardiol*. 2012;9(9):526–539. doi:10.1038/nrcardio.2012.91
16. Avram R, Tison GH, Aschbacher K, et al. Real-world heart rate norms in the Health eHeart study. *npj Digital Medicine*. 2019;2(1). doi:10.1038/s41746-019-0134-9
17. Mancia G. EFFECTS OF BLOOD-PRESSURE MEASUREMENT BY THE DOCTOR ON PATIENT'S BLOOD PRESSURE AND HEART RATE. *The Lancet*. 1983;322(8352):695-698. doi:10.1016/S0140-6736(83)92244-4
18. Pierdomenico SD, Bucci A, Lapenna D, Cuccurullo F, Mezzetti A. Clinic and ambulatory heart rate in sustained and white-coat hypertension: *Blood Pressure Monitoring*. 2001;6(5):239-244. doi:10.1097/00126097-200110000-00003
19. Geenen R, Vijver FJR. A simple test of the Law of Initial Values. *Psychophysiology*. 1993;30(5):525-530. doi:10.1111/j.1469-8986.1993.tb02076.x
20. Pattyn N, Migeotte P-F, Neyt X, den Nest A van, Cluydts R. Comparing real-life and laboratory-induced stress reactivity on cardio-respiratory parameters: Differentiation of a tonic and a phasic component. *Physiology & Behavior*. 2010;101(2):218-223. doi:10.1016/j.physbeh.2010.04.037
21. Wilder J. BASIMETRIC APPROACH (LAW OF INITIAL VALUE) TO BIOLOGICAL RHYTHMS. *Annals of the New York Academy of Sciences*. 1962;98(4):1211-1220. doi:10.1111/j.1749-6632.1962.tb30629.x
22. Rohleder N. Stress and inflammation – The need to address the gap in the transition between acute and chronic stress effects. *Psychoneuroendocrinology*. February 2019. doi:10.1016/j.psyneuen.2019.02.021

23. Mahloko L, Adebessin F. A Systematic Literature Review of the Factors that Influence the Accuracy of Consumer Wearable Health Device Data. In: Hattingh M, Matthee M, Smuts H, Pappas I, Dwivedi YK, Mäntymäki M, eds. *Responsible Design, Implementation and Use of Information and Communication Technology*. Vol 12067. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2020:96-107. doi:10.1007/978-3-030-45002-1\_9
24. Elgendi M. On the Analysis of Fingertip Photoplethysmogram Signals. *Current Cardiology Reviews*. 2012;8(1):14-25. doi:10.2174/157340312801215782
25. Bradshaw B, Konty K, Ramirez E, Lee W, Signorini A, Foschini L. Influenza Surveillance Using Wearable Mobile Health Devices. In: ; 2019.
26. Natarajan A, Pantelopoulos A, Emir-Farinas H, Natarajan P. Heart Rate Variability with Photoplethysmography in 8 Million Individuals: Results and Scaling Relations with Age, Gender, and Time of Day. *bioRxiv*. September 2019. doi:10.1101/772285
27. Perez MV, Mahaffey KW, Hedlin H, et al. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. *New England Journal of Medicine*. 2019;381(20):1909-1917. doi:10.1056/NEJMoa1901183
28. Sim I. Mobile Devices and Health. *New England Journal of Medicine*. 2019;381(10):956-968. doi:10.1056/NEJMr1806949
29. Tison GH, Sanchez JM, Ballinger B, et al. Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch. *JAMA Cardiology*. 2018;3(5):409. doi:10.1001/jamacardio.2018.0136
30. Tison GH, Singh AC, Ohashi DA, et al. Cardiovascular Risk Stratification Using Off-the-Shelf Wearables and a Multi-Task Deep Learning Algorithm. In: ; 2018.
31. Stuppel A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. *npj Digital Medicine*. 2019;2(1). doi:10.1038/s41746-019-0079-z
32. Nelson BW, Allen NB. Extending the Passive-Sensing Toolbox: Using Smart-Home Technology in Psychological Science. *Perspectives on Psychological Science*. 2018;13(6):718-733. doi:10.1177/1745691618776008
33. Hunkin H, King DL, Zajac IT. Wearable devices as adjuncts in the treatment of anxiety-related symptoms: A narrative review of five device modalities and implications for clinical practice. *Clinical Psychology: Science and Practice*. 2019;26(3). doi:10.1111/cpsp.12290
34. Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. *The Lancet Digital Health*. January 2020. doi:10.1016/S2589-7500(19)30222-5

35. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. *International Journal of Behavioral Nutrition and Physical Activity*. 2015;12(1). doi:10.1186/s12966-015-0314-1
36. 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. *Frontiers in Physiology*. 2018;9. doi:10.3389/fphys.2018.00743
37. Bland JM, Altman DG. *Measuring Agreement in Method Comparison Studies*. Vol 8.; 1999. doi:10.1191/096228099673819272
38. Nelson BW, Allen NB. Accuracy of Consumer Wearable Heart Rate Measurement During an Ecologically Valid 24-Hour Period: Intraindividual Validation Study. *JMIR mHealth and uHealth*. 2019;7(3):e10828. doi:10.2196/10828
39. Boudreaux BD, Hebert EP, Hollander DB, et al. Validity of Wearable Activity Monitors during Cycling and Resistance Exercise: *Medicine & Science in Sports & Exercise*. 2018;50(3):624-633. doi:10.1249/MSS.0000000000001471
40. de Zambotti M, Baker FC, Willoughby AR, et al. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiology & Behavior*. 2016;158:143-149. doi:10.1016/j.physbeh.2016.03.006
41. Kroll RR, Boyd JG, Maslove DM. Accuracy of a Wrist-Worn Wearable Device for Monitoring Heart Rates in Hospital Inpatients: A Prospective Observational Study. *Journal of Medical Internet Research*. 2016;18(9):e253. doi:10.2196/jmir.6025
42. Shcherbina A, Mattsson C, Waggott D, et al. Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *Journal of Personalized Medicine*. 2017;7(2):3. doi:10.3390/jpm7020003
43. Wallen MP, Gomersall SR, Keating SE, Wisløff U, Coombes JS. Accuracy of heart rate watches: Implications for weight management. *PLoS ONE*. 2016;11(5). doi:10.1371/journal.pone.0154420
44. Wang R, Blackburn G, Desai M, et al. Accuracy of wrist-worn heart rate monitors. *JAMA Cardiology*. 2017;2(1):104-106. doi:10.1001/jamacardio.2016.3340
45. Dooley EE, Golaszewski NM, Bartholomew JB. Estimating Accuracy at Exercise Intensities: A Comparative Study of Self-Monitoring Heart Rate and Physical Activity Wearable Devices. *JMIR mHealth and uHealth*. 2017;5(3):e34. doi:10.2196/mhealth.7043
46. Gillinov S, Etiwy M, Wang R, et al. Variable accuracy of wearable heart rate monitors during aerobic exercise. *Medicine and Science in Sports and Exercise*. 2017;49(8):1697-1703. doi:10.1249/MSS.0000000000001284

47. Stahl SE, An H-S, Dinkel DM, Noble JM, Lee J-M. How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport & Exercise Medicine*. 2016;2(1):e000106. doi:10.1136/bmjsem-2015-000106
48. El-Amrawy F, Nounou MI. Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial? *Healthcare Informatics Research*. 2015;21(4):315–320. doi:10.4258/hir.2015.21.4.315
49. FDA. Electrocardiograph software for over-the-counter use. 2018. [https://www.accessdata.fda.gov/cdrh\\_docs/pdf18/DEN180044.pdf](https://www.accessdata.fda.gov/cdrh_docs/pdf18/DEN180044.pdf).
50. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. *npj Digital Medicine*. 2020;3(1). doi:10.1038/s41746-020-0226-6
51. Xie J, Wen D, Liang L, Jia Y, Gao L, Lei J. Evaluating the Validity of Current Mainstream Wearable Devices in Fitness Tracking Under Various Physical Activities: Comparative Study. *JMIR Mhealth Uhealth* 2018;6(4):e94 <http://mhealth.jmir.org/2018/4/e94/>. 2018;6(4):e94. doi:10.2196/MHEALTH.9754
52. Benedetto S, Caldato C, Bazzan E, Greenwood DC, Pensabene V, Actis P. Assessment of the fitbit charge 2 for monitoring heart rate. *PLoS ONE*. 2018. doi:10.1371/journal.pone.0192691
53. Bavel JJV, Mende-siedlecki P, Brady WJ, Reinero DA. Contextual sensitivity in scientific reproducibility Submission PDF Contextual sensitivity in scientific reproducibility. 2016;I(1):6454–6459. doi:10.1073/pnas.1521897113
54. Munafò MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. *Nature Human Behaviour*. 2017;1(1). doi:10.1038/s41562-016-0021
55. Tackett JL, Brandes CM, King KM, Markon KE. Psychology's Replication Crisis and Clinical Psychological Science. *Annual Review of Clinical Psychology*. 2019;15(1):579-604. doi:10.1146/annurev-clinpsy-050718-095710
56. Hastings, Jr. C, Mosteller, F, Tukey JW, Winsor CP. Low Moments for Small Samples: A Comparative Study of Order Statistics. *Ann Math Statist*. 1947;18(3):413-426.
57. Yang Z, Zhou Q, Lei L, Zheng K, Xiang W. An IoT-cloud Based Wearable ECG Monitoring System for Smart Healthcare. *Journal of Medical Systems*. 2016;40(12). doi:10.1007/s10916-016-0644-9
58. Graham JW. Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*. 2009;60(1):549-576. doi:10.1146/annurev.psych.58.110405.085530
59. Ji L, Chow S-M, Schermerhorn AC, Jacobson NC, Cummings EM. Handling Missing Data in the Modeling of Intensive Longitudinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*. 2018;25(5):715-736. doi:10.1080/10705511.2017.1417046

60. Jacobson NC, Chow S-M, Newman MG. The Differential Time-Varying Effect Model (DTVEM): A tool for diagnosing and modeling time lags in intensive longitudinal data. *Behavior Research Methods*. 2019;51(1):295-315. doi:10.3758/s13428-018-1101-0
61. Gorny AW, Liew SJ, Tan CS, Müller-Riemenschneider F. Fitbit Charge HR Wireless Heart Rate Monitor: Validation Study Conducted Under Free-Living Conditions. *JMIR mHealth and uHealth*. 2017. doi:10.2196/mhealth.8233
62. Menghini L, Gianfranchi E, Cellini N, Patron E, Tagliabue M, Sarlo M. Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiology*. 2019;56(11). doi:10.1111/psyp.13441
63. Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: Unlocking the file drawer. *Science*. 2014;345(6203):1502-1505. doi:10.1126/science.1255484
64. Button KS, Ioannidis JPA, Mokrysz C, et al. Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience*. 2013;14(8):585-585. doi:10.1038/nrn3475-c4
65. Nowok B, Raab G, Dibben C. *Synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control*.; 2016. <https://cran.r-project.org/web/packages/synthpop/index.html>.
66. Tell it like it is. *Nature Human Behaviour*. 2020;4(1):1-1. doi:10.1038/s41562-020-0818-9
67. Weiler DT, Villajuan SO, Edkins L, Cleary S, Saleem JJ. Wearable Heart Rate Monitor Technology Accuracy in Research: A Comparative Study Between PPG and ECG Technology. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2017;61(1):1292-1296. doi:10.1177/1541931213601804
68. Hall GCN, Yip T, Zárata MA. On becoming multicultural in a monocultural research world: A conceptual approach to studying ethnocultural diversity. *American Psychologist*. 2016;71(1):40-51. doi:10.1037/a0039734
69. Cosoli G, Spinsante S, Scalise L. Wrist-worn and chest-strap wearable devices: Systematic review on accuracy and metrological characteristics. *Measurement*. 2020;159:107789. doi:10.1016/j.measurement.2020.107789
70. Bai Y, Hibbing P, Mantis C, Welk GJ. Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR. *Journal of Sports Sciences*. 2018. doi:10.1080/02640414.2017.1412235
71. Cadmus-Bertram L, Gangnon R, Wirkus EJ, Thraen-Borowski KM, Gorzelitz-Liebhauser J. The accuracy of heart rate monitoring by some wrist-worn activity trackers. *Annals of internal medicine*. 2017;166(8):610–612.

72. Jo E, Lewis K, Directo D, Kim MJY, Dolezal BA. Validation of biofeedback wearables for photoplethysmographic heart rate tracking. *Journal of Sports Science and Medicine*. 2016;15(3):540–547.
73. Spierer DK, Rosen Z, Litman LL, Fujii K. Validation of photoplethysmography as a method to detect heart rate during rest and exercise. *Journal of Medical Engineering and Technology*. 2015. doi:10.3109/03091902.2015.1047536
74. Parak J, Korhonen I. Evaluation of wearable consumer heart rate monitors based on photoplethysmography. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*. 2014:3670–3673. doi:10.1109/EMBC.2014.6944419
75. Valencell T. Valencell | Optical Heart Rate Monitoring: What You Need to Know. Valencell. <https://valencell.com/blog/2015/10/optical-heart-rate-monitoring-what-you-need-to-know/>. Published October 15, 2015. Accessed April 13, 2020.
76. O'Connor MF, Bower JE, Cho HJ, et al. To assess, to control, to exclude: Effects of biobehavioral factors on circulating inflammatory markers. *Brain, Behavior, and Immunity*. 2009;23(7):887–897. doi:10.1016/j.bbi.2009.04.005
77. Quintana DS, Alvares GA, Heathers JAJ. Guidelines for Reporting Articles on Psychiatry and Heart rate variability (GRAPH): recommendations to advance research communication. *Translational Psychiatry*. 2016;6(5):e803-e803. doi:10.1038/tp.2016.73
78. IDC. Shipments of Wearable Devices Reach 118.9 Million Units in the Fourth Quarter and 336.5 Million for 2019, According to IDC. IDC: The premier global market intelligence company. <https://www.idc.com/getdoc.jsp?containerId=prUS46122120>. Published 2020. Accessed March 31, 2020.
79. Apple. Your heart rate. What it means, and where on Apple Watch you'll find it. Apple Support. <https://support.apple.com/en-gb/HT204666>. Published 2020. Accessed March 31, 2020.
80. Fitbit. Fitbit PurePulse® Continuous Wrist-Based Heart Rate. <https://www.fitbit.com/eu/purepulse>. Published 2020. Accessed March 31, 2020.
81. Garmin. Garmin Elevate Optical Heart Rate | Garmin. <https://www.garmin.com.sg/garmin-technology/heart-rate>. Published 2020. Accessed March 31, 2020.
82. Huawei. Measuring heart rate. <https://consumer.huawei.com/en/support/content/en-us00756088/>. Published 2020. Accessed April 16, 2020.

## **Competing Interests**

Dr. Jacobson is the owner of a free application published on the Google Play Store entitled “Mood Triggers”. He does not receive any direct or indirect revenue from his ownership of the application (i.e. the application is free, there are no advertisements, and the data is only being used for research purposes). Dr. Jacobson has no other disclosures to report. All other authors declare that there are no competing interests.



## **Author Contribution**

Benjamin W. Nelson, M.S.

- Provided conception of manuscript content and design, organized the collaboration, and drafted manuscript.

Carissa A. Low, Ph.D.

- Assisted in manuscript drafting and revision for critically important intellectual content.

Nicholas Jacobson, Ph.D.

- Assisted in manuscript drafting and revision for critically important intellectual content and provided methodological expertise.

Patricia Areán, Ph.D.

- Assisted in manuscript drafting and revision for critically important intellectual content.

John Torous, M.D.

- Assisted in manuscript drafting and revision for critically important intellectual content.

Nicholas B. Allen, Ph.D.

- Assisted in manuscript drafting and revision for critically important intellectual content and provided psychophysiological expertise.