Title: Accuracy of Wearables During a Continuous and Ecologically Valid 24-Hour Period That Approximates Actual Consumer Device Use Conditions Within an Individual.

Authors: Benjamin W. Nelson, M.S.\textsuperscript{a,b} & Nicholas B. Allen, Ph.D.\textsuperscript{a,b}

Affiliations:
\begin{itemize}
  \item a. Department of Psychology, University of Oregon, Eugene, OR, USA
  \item b. The Center for Digital Mental Health (CDMH), University of Oregon, Eugene, OR, USA
\end{itemize}

Corresponding Author:

Benjamin W. Nelson

Department of Psychology

1227 University of Oregon

Eugene, OR 97403 USA

Email: bwn@uoregon.edu
Abstract

Background: Wrist-worn smart watches and fitness monitors (i.e., wearables) have become widely adopted by consumers and are gaining increased attention from researchers for their potential contribution to digital measurement of health in a scalable, and ecologically valid way. Various studies have begun to examine the accuracy of these devices in controlled laboratory settings (e.g., treadmill, stationary bike), yet no studies to date have investigated the accuracy of wearables during a continuous and ecologically valid 24-hour period that approximates actual consumer device use conditions.

Objective: The goal of the present study was to determine the accuracy of two popular wearable devices, the Apple Watch 3 and Fitbit Charge 2, to assess heart rate accuracy as compared to the gold-standard reference method for real-world settings, an ambulatory electrocardiogram (ECG). Data was collected across five daily conditions, including sitting, walking, running, daily activities (e.g., cleaning, chores, work transitions), and sleeping.

Methods: One participant, (BN; 29-year-old Caucasian male) completed a 24-hour ecologically valid protocol by wearing two of the most popular wrist wearable devices (Apple Watch 3 and Fitbit Charge 2). In addition, an ambulatory ECG (Vrije Universiteit Ambulatory Monitoring System) was used as the reference method during a normal day, which resulted in the collection of 102,740 individual heart beats. A single subject design was used to keep all variables constant except for wearable devices. Accuracy of these devices compared to the gold-standard ECG was assessed using percent error, Bland-Altman analyses, and concordance class correlation to assess agreement between devices.
Results: The Apple Watch 3 and Fitbit Charge 2 were generally highly accurate across the overall 24-hour condition. Specifically, the Apple Watch 3 had a mean difference of -1.80 bpm, a mean error percent of -2.25%, and a mean agreement of 95% when compared to the ECG across 24 hours. The Fitbit Charge 2 had a mean difference of -3.47 bpm, a mean error of -4.25%, and a mean agreement of 91% when compared to the ECG across 24 hours. These findings varied by condition.

Conclusion: The Apple Watch 3 and the Fitbit Charge 2 provided acceptable heart rate accuracy overall across a 24 hour continuous and ecologically valid period as compared to the gold-standard ECG, although during some conditions and devices error rose past the 5% acceptable error rate and some individual measurements were significantly erratic. Overall, these devices appear to be useful for implementing as ambulatory measures of cardiac activity in research studies, especially those where the specific advantages of these methods (e.g., scalability, low participant burden) are particularly suited to the population or research question.

Keywords: ambulatory electrocardiogram; Apple Watch 3; digital sensors; Fitbit Charge 2; passive sensing; photoplethysmography; wearables
Wrist-worn smart watches and fitness monitors or wearables have been widely adopted by consumers and are currently gaining increased attention by researchers for their potential contribution to digital measurement of health, especially in “big data” studies as they are scalable, unobtrusive, and provide greater ecologically validity. These devices contain a multitude of sensors, including an optical sensor that uses photoplethysmography (PPG), allowing these devices to collect heart rate (HR). Recently, there have been a variety of studies that have examined the accuracy of wearable heart rate sensors as compared to an electrocardiogram (ECG) [1-7], Polar chest strap [8,9], or pulse oximeter [10] across various controlled laboratory conditions including sitting, treadmill protocols for walking and running, cycling, weight training, and sleeping. However, to our knowledge no studies have tested the accuracy of HR sensors on these devices as they were devised to be used by consumers - out in the real world during a 24-hour ecologically valid paradigm that approximates actual consumer device use conditions.

Previous research comparing wearables to the gold-standard ECG have shown that that wearables underestimate absolute HR as compared to reference methods [1,2,4,6-9]. Prior research has also shown that the Apple Watch has greater accuracy than Fitbit devices [5-7]. Specifically, prior research has found that the Apple Watch has lower overall error [1,5,8], lowest mean difference standard deviation [6], and higher agreement with ECG than Fitbit devices [1,7], but that wearables accuracy depends on activity [3]. Specifically, research shows that position of wearable device impacts accuracy [11] and that wearable devices are more accurate during rest and low intensity exercise as compared to exercises at higher intensity [1,12], which may be due to less movement of the wearable around the wrist, although this is not found in all studies [8].
The Current Study:

This study was pre-registered (hypotheses, methods) with open code and data on Open Science Framework (see osf.io/6w2sh). The objective of this study was to determine the HR accuracy of two of the most popular wearables, the Apple Watch 3 and Fitbit Charge 2, as compared to the gold-standard ECG. We used an ambulatory ECG to allow for continuous recording in real-world settings. A single-subject design was used for this initial study on the ecological validity of wearables, because it allowed for many potential confound variables to be held constant, except for the wearable devices, thus providing a powerful test of the accuracy of the devices per se.

The current study hypothesized that 1) the Apple Watch 3 would be more accurate at measuring HR than the Fitbit Charge 2 when compared to an ambulatory ECG across all conditions, 2) both wearables would underestimate HR across all conditions, and 3) that device measurement of HR would become increasingly inaccurate as activity intensity increased.

Methods:

Participant:

This study contained one subject (BN) who completed a 24-hour protocol (29-year-old Caucasian male; Body Mass Index = 21.1; Fitzpatrick skin tone measure = 2; Right Wrist (cm) = 7.0; Left Wrist (cm) = 6.5; Right Hand Dominant). The participant (1st author) conceptualized and initiated this study with the purpose of having the data published. Therefore, approval from the University of Oregon ethics committee was unnecessary and not obtained. The participant gave consent for collecting and using the data for study purposes.

Study Protocol:
Participant psychophysiology recordings began at 18:28 on Day 1 and briefly stopped at 17:10 on Day 2 prior to the run condition. Recording resumed at 17:37 for the run condition and stopped at 18:50 on Day 2. Age, gender, height, and weight were used to set up both wearable devices.

Conditions:

Five daily conditions were recorded throughout the 24-hour study using a digital notebook (Google Sheets) to record activity times, resulting in 84 start and stop marker times. These included sitting, which included any seated activity; walking; running (this occurred on a treadmill to allow for a stable ambulatory ECG to increase accuracy); daily activities, which included activities such as cleaning, vacuuming, and cooking; and sleeping.

Gold-Standard Reference Method:

Electrocardiography (ECG): ECG data were acquired using a standard 3-lead ambulatory ECG (Vrije Universiteit Ambulatory Monitoring System) [13,14].

Wearable Devices:

Apple Watch 3: The Apple Watch Series 3 (2017 version, Apple Inc, California, USA, v. 4.2.3) 42mm was worn on the right wrist. According to Apple, the Apple Watch 3 samples HR approximately every 10 minutes or continuously during workouts using PPG with either green LED or infrared light and photodiode sensors. All data from the Apple Watch 3 was sync with the Apple Health app on the iPhone and then exported in XML format for analysis. The AppleHealthAnalysis GitHub repository [15] was used to convert the XML file to a dataframe in R Studio to access per minute data for analysis. When more than one heart rate measurement was
collected each minute, the average of these measurements was used in line with prior wearable research [5].

Fitbit Charge 2: The Fitbit Charge 2 (2017 version, Fitbit Inc, California, USA, v. 22.55.2) was worn on the left wrist. According to Fitbit, the Fitbit Charge 2 samples HR at varying rates depending on activity level using PPG. The fitbitr GitHub repository [16] was used to interact with the Fitbit application programing interface (API) to access per minute data for analysis.

Error: In line with prior health sciences research on wearable HR accuracy [5] and pedometer step counting accuracy [17] we defined an acceptable error rate to be ± 5%. Outliers were not removed as this would interfere with accuracy statistics.

Statistical Analysis:

All analyses were performed in R (version 3.4.3) using RStudio (version 1.1.383). Scripts can be found on Github and OSF and data can be found on OSF. Analyses were performed using the average beats per minute (bpm) separately for each wearable device. ECG data was used as the gold-standard for HR calculated as bpm.

Percent Error. The percent error relative to the ECG was calculated for heart rate in line with previous wearable research [5] for each wearable by using the following formula:

\[
Percent Error = \left( \frac{\text{device measurement} - \text{gold standard}}{\text{gold standard}} \right) \times 100 .
\]
Bland-Altman Analysis: Bland-Altman Analysis and 95% Limits of Agreement were calculated using the BlandAltmanLeh R package [18] for the main analyses, rather than concordance class correlation, to determine agreement between devices as this is the main method used for comparing medical instruments [19,20] and research indicates that different methods are unlikely to have exact agreement and therefore the importance lies in how close pairs of observations are as small differences between devices are unlikely to impact patient decisions [21].

Concordance Class Correlation (CCC): Lastly, although not one of the analyses that was pre-registered, we also ran CCC analyses between the ECG and each wearable device separately across all conditions using the DescTools R Package [22] to assist in Bland-Altman Plot Interpretation. In line with prior wearable research [6], the strength of agreement was interpreted based on the following, weak (ccc < 0.5), moderate (ccc = 0.5-0.7), and strong (ccc > 0.7).

Results:

Descriptives: The ECG collected 102,740 individual heart beat recordings, which resulted in 1,424 individual beat per minute (bpm) observations after data cleaning and then being condensed to bpm observations. The Fitbit Charge 2 collected 1,446 bpm observations and the Apple Watch 3 collected 1,545 individual observations, which resulted in 394 bpm observations after averaging multiple observations within the same minute as described above and in line with prior wearable research. Overall, 4,415 raw bpm observations or 3,264 cleaned bpm observations (averaging multiple Apple Watch 3 observations within a single minute), which is up to 84% more data within a single subject as compared to some prior studies that had 50 subjects (Wang
et al., 2016). See Table 1 for number of observations, HR descriptives, mean error percentage, Bland-Altman Analyses, and CCC agreement for each condition.
<table>
<thead>
<tr>
<th>Activity</th>
<th>Device</th>
<th>Number of Observations</th>
<th>HR Descriptives</th>
<th>Percent Error</th>
<th>Bland-Altman Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>HR Mean (SD)</td>
<td>HR Range</td>
<td>HR Mean Percentage (SD)</td>
</tr>
<tr>
<td>24 Hours</td>
<td>ECG</td>
<td>1424</td>
<td>72.65 (16.92)</td>
<td>51-161</td>
<td>-2.25% -1.80 -12.71</td>
</tr>
<tr>
<td></td>
<td>Apple</td>
<td>394</td>
<td>78.78 (25.74)</td>
<td>49-165</td>
<td>-4.25% -3.47 8.62</td>
</tr>
<tr>
<td></td>
<td>Fitbit</td>
<td>1446</td>
<td>69.10 (15.10)</td>
<td>50-153</td>
<td>-6.29% -4.69 4.91</td>
</tr>
<tr>
<td></td>
<td>Charg</td>
<td>535</td>
<td>70.41 (7.24)</td>
<td>55-97</td>
<td>3.14% 2.47 12.01</td>
</tr>
<tr>
<td>Sitting</td>
<td>ECG</td>
<td>100</td>
<td>102.32 (16.87)</td>
<td>61-127</td>
<td>.14% .11 (7.29) 14.41</td>
</tr>
<tr>
<td></td>
<td>Apple</td>
<td>79</td>
<td>106.06 (15.03)</td>
<td>55-139</td>
<td>-6.50% -6.85 14.81</td>
</tr>
<tr>
<td></td>
<td>Fitbit</td>
<td>100</td>
<td>95.47 (17.88)</td>
<td>54-132</td>
<td>10.20 (11.05) 28.51</td>
</tr>
<tr>
<td>Walking</td>
<td>ECG</td>
<td>22</td>
<td>147.82 (19.48)</td>
<td>104-161</td>
<td>-6.85% -6.85 14.81</td>
</tr>
<tr>
<td>Running</td>
<td>ECG</td>
<td>22</td>
<td>147.82 (19.48)</td>
<td>104-161</td>
<td>-6.85% -6.85 14.81</td>
</tr>
</tbody>
</table>
Percent Error:

Overall across the 24-hour recording, the Apple Watch 3 had a mean percent error of -2.25%, while the Fitbit Charge 2 had a mean percent error of -4.25%. During sitting conditions, the Apple Watch 3 had a mean percent error of -3.14%, while the Fitbit Charge 2 had a mean percent error of -6.29%. During walking conditions, the Apple Watch 3 had a mean percent error of .14%, while the Fitbit Charge 2 had a mean percent error of -6.50%. During the running condition, the Apple Watch 3 had a mean percent error of 1.50%, while the Fitbit Charge 2 had a mean percent error of -9.88%. During daily activity conditions, the Apple Watch 3 had a mean percent error of -9.38%, while the Fitbit Charge 2 had a mean percent error of -4.33%. Lastly, during the sleep condition, the Apple Watch 3 had a mean percent error of -1.36%, while the Fitbit Charge 2 had a mean percent error of -1.62%.

Bland-Altman Analysis and 95% Limits of Agreement:

Overall, across the 24-hour recording (see Figure 1a and 1b) the Apple Watch 3 had a mean error of -1.80 bpm (Lower LoA-Upper LoA; -16.31 to 12.71 bpm), while the Fitbit Charge 2 had a mean error of -3.47 bpm (Lower LoA-Upper LoA; -15.54 to 8.62 bpm). Visual inspection of the Bland-Altman plots (see Figure 1a and 1b) revealed a tendency for the Apple Watch 3 to both over and underestimate HR values when HR values were between 70-120 bpm, while the Fitbit Charge 2 had a tendency to underestimated HR values, particularly once HR values exceeded ~80 bpm.
During sitting conditions (see Figure 2a and 2b), the Apple Watch 3 had a mean error of -2.47 bpm (Lower LoA-Upper LoA: -16.94 to 12.01 bpm), while the Fitbit Charge 2 had a mean error of -4.69 bpm (Lower LoA-Upper LoA: -14.29 to 4.91 bpm).

During walking conditions (see Figure 2c and 2d), the Apple Watch 3 had a mean error of 0.11 bpm (Lower LoA-Upper LoA: -14.18 to 14.41 bpm), while the Fitbit Charge 2 had a mean error of -6.85 bpm (Lower LoA-Upper LoA: -28.51 to 14.81 bpm).
During the running condition (see Figure 2e and 2f), the Apple Watch 3 had a mean error of 1.77 bpm (Lower LoA-Upper LoA; 9.78 to 13.33 bpm), while the Fitbit Charge 2 had a mean error of -14.73 bpm (Lower LoA-Upper LoA; -29.77 to 0.31 bpm).

During daily activity conditions (see Figure 2g and 2h), the Apple Watch 3 had a mean error of -8.50 bpm (Lower LoA-Upper LoA; -33.78 to 16.78 bpm), while the Fitbit Charge 2 had a mean error of -3.73 bpm (Lower LoA-Upper LoA; -19.88 to 12.41 bpm).
Lastly, during the sleep condition (see Figure 2i and 2j), the Apple Watch 3 had a mean error of -0.95 bpm (Lower LoA-Upper LoA; -6.39 to 4.50 bpm), while the Fitbit Charge 2 had a mean error of -1.11 bpm (Lower LoA-Upper LoA; -7.28 to 5.17 bpm).

Concordance Class Correlation (CCC):

Overall, across the 24-hour recording the Apple Watch 3 (ccc = .955, 95% CI [.945, .963]) and the Fitbit Charge 2 (ccc = .906, 95% CI [.896, .914]) had strong agreement with the
reference method. During sitting conditions, the Apple Watch 3 (ccc = .453, 95% CI [.321, .567]) had weak agreement and the Fitbit Charge 2 (ccc = .561, 95% CI [.515, .603]) had moderate agreement with the reference method. During all walking activities, the walking conditions the Apple Watch 3 (ccc = .871, 95% CI [.807, .915]) and the Fitbit Charge 2 (ccc = .740, 95% CI [.645, .812]) had strong agreement with the reference method. During the running condition, the Apple Watch 3 (ccc = .864, 95% CI [.731, .934]) had strong agreement with the reference method, while the Fitbit Charge 2 (ccc = .490, 95% CI [.268, .663]) had weak agreement with the reference method. During all daily activity conditions, the Apple Watch 3 (ccc = .460, 95% CI [.204, .656]) had weak agreement with the reference method, while the Fitbit Charge 2 (ccc = .739, 95% CI [.676, .791]) had strong agreement with the reference method. Lastly during the sleep condition, the Apple Watch 3 (ccc = .791, 95% CI [.715, .849]) and the Fitbit Charge 2 (ccc = .745, 95% CI [.707, .779]) had strong agreement with the reference method.

Discussion:

This study provided the first continuous and ecologically valid assessment of the accuracy of the Apple Watch 3 and the Fitbit Charge 2 as they were devised to be used by consumers (i.e., during ecologically valid daily activities) during a 24-hour paradigm that approximated actual consumer device use conditions.

In line with previous controlled laboratory research [2,3,5-7,9,10], our findings indicated that both wearable devices provided acceptable accuracy overall across the 24-hour recording period. In addition, in line with previous research both the Apple Watch 3 and the Fitbit Charge 2 slightly underestimated heart rate as compared to ECG and other reference methods [1,2,4,6-9]. Although this overall (across the 24-hour study) underreporting of absolute HR is unlikely to be
problematic in most contexts (< 4bpm), particularly those where the primary interest is in the pattern of change in HR associated with different conditions or activities, there were a number of individual observations that were inaccurate by significantly large margins, which would be problematic in some contexts (e.g., medical settings). This indicates that while overall, summary statistics may be very accurate for research purposes, any single observation in real-time may have a large degree of error. In addition, we found it surprising that the Apple Watch 3 had such a high mean percent error rate (-9.38%) during daily activities as compared to the Fitbit Charge 2 mean percent error (4.33%). This difference may be due to the fact that the Apple Watch 3 was worn on the dominant hand, which may have made more erratic movements than the Fitbit Charge 2 on the non-dominant hand during daily activities potentially moving the position of the wearable and making it more difficult for the PPG sensor to assess and accurate HR measurement [11].

Overall, the Apple Watch 3 had acceptable error across the entire 24-hour period as well as during the sitting, walking, running, and sleeping conditions, while it’s error rate rose above the ± 5% threshold for daily activities (-9.38%). In addition, the Fitbit Charge 2 had acceptable error across the entire 24-hour period as well as during the sitting, daily activity, and sleeping, but rose above the ± 5% threshold for walking (-6.50%) and running (-9.88%).

The current study had a number of strengths. First, the time intensive single-subject design allowed all potential confounding variables to be constant except for the wearable devices. Second, the length of recording resulted in the collection of a total of 3,264 bpm observations, which is up to 84% more data within a single subject as compared to some prior study data across 50 subjects [7]. Lastly, this study provided the first continuous and ecologically valid assessment of wearable HR accuracy in real-world conditions. In addition to these
strengths, there were also a number of limitations. First, the single-subject design limited various participant demographic factors, such as higher BMI, darker skin tone, and larger wrist circumference, which have been shown to positively correlate with HR error rates [5]. Future studies should attempt to replicate these results across multiple individuals with diverse BMI, wrist circumference, skin tone, fitness level, and stress level. In addition, the single subject design combined with the Apple Watch 3 sampling rate of approximately every 10 minutes led to a small number of observations for some conditions. While continuous recording was not activated on the Apple Watch 3 in order to approximate real-world usage conditions, future studies should aim to collect larger numbers of subjects in order to increase the observations for each condition and potentially activate continuous recording on this device. Similarly, while this study had the strength of providing the first continuous and ecologically valid assessment of wearable accuracy in real-world conditions this was also a limitation as it inherently couldn’t take place within more controlled laboratory settings that used a stationary ECG, rather than an ambulatory ECG that may introduce some additional error. Another limitation to this study is that while that overall error rate of both devices was low, there were some individual observations that were inaccurate by significantly large margins. This indicates that while overall, summary statistics for conditions may be very accurate, any single observation in real-time may have a large degree of error. Researchers should keep this in mind when using wearable devices in research settings and this finding emphasizes the importance of data cleaning. Implementing these devices in research settings would likely benefit from automated outlier detection and deletion techniques. Lastly, this study did not counterbalance wrist placement of the wearables to rule out potential influences of wrist circumference or musculature on the accuracy of HR readings. The subject was right handed, and therefore the lower accuracy of the Apple Watch 3 as
compared to the Fitbit Charge 2 during the daily activities condition may have been due to inconsistent wrist motions that accompany many activities in this condition as prior research has indicated that the lack of smooth wrist movements introduces larger HR measurement error [8]. Future studies should provide both between-subjects analyses and within-subjects analyses with devices on both wrists to assess the accuracy of wearables as hand dominance may influence accuracy.

This study provided the first continuous and ecologically valid assessment of the accuracy of the Apple Watch 3 and the Fitbit Charge 2 HR measurements as they were devised to be used by consumers out in the real world during a 24-hour paradigm that approximated actual consumer device use conditions. Overall, both the Apple Watch 3 and Fitbit Charge 2 had acceptable HR accuracy overall across the 24-hour period with the Apple Watch 3 having acceptable HR error across the day as well as the during the sitting, walking, running, and sleeping conditions, while the Fitbit Charge 2 had acceptable HR error across the entire day as well as during the sitting, daily activity, and sleeping. In contrast, the Apple Watch did not have acceptable accuracy during daily activities, while the Fitbit Charge 2 did not have acceptable accuracy during walking and running. Again it is important to note that while overall statistics for most conditions were acceptable, there were a number of individual observations that varied widely from the gold-standard ECG, which indicates that any single measurement viewed in real-time cannot be interpreted as an accurate measurement. Overall, wearable devices likely won’t be replacing the gold-standard ECG in a medical setting anytime soon, but both the Apple Watch 3 and the Fitbit Charge 2 are acceptable for research and clinical applications, particularly big data studies, as these devices had an overall acceptable error rate combined with being relatively cheap, unobtrusive, and scalable as compared to gold-standard medical equipment.
References:


