2019

Validity of Heart Rate Measurements in a Fitbit Charge 2 and Apple iWatch in a Free-Living Setting

Dakota R. Dione

Recommended Citation
Dione, Dakota R., "Validity of Heart Rate Measurements in a Fitbit Charge 2 and Apple iWatch in a Free-Living Setting" (2019). *UVM Honors College Senior Theses*. 281.
https://scholarworks.uvm.edu/hcoltheses/281

VALIDITY OF HEART RATE MEASUREMENTS IN THE FITBIT CHARGE 2 AND APPLE IWATCH IN A SHORT-TERM FREE-LIVING SETTING

By Dakota Dione
Honors College Senior Thesis
Department of Rehabilitation and Movement Science
College of Nursing and Health Sciences
University of Vermont
May 2019

ABSTRACT

The purpose of this investigation was to investigate the validity of heart rate (HR) measurements in Fitbit Charge 2 (FBC2) and Apple iWatch (AW). Fifty-two healthy adults wore 7 activity monitors while engaging in a normal daily activity free-living data collection for a twenty-four-hour period. HR criterion measures were collected by a Polar H7 HR chest strap monitor. For estimating whole-HR validity, the mean absolute percent errors were smaller in magnitude for sedentary behavior, moderate physical activity, and vigorous physical activity for the FBC2 (4%, 10%, 14%) compared to the AW (7%, 10%, 16%). Bland-Altman analysis revealed both FBC2 and AW tended to underestimate HR measurement values. FBC2 was considered in agreement for sedentary behavior and moderate intensity physical activity. To a lesser degree, AW was considered in agreement for sedentary behavior and moderate intensity physical activity. Both FBC2 and AW had a weak-moderate correlation for vigorous intensity physical activity (r_{FBC2}=0.49, p= <0.0001 and r_{AW}=0.49, p=<0.0001). The FBC2 and AW should be used with caution, however they have reasonable validity for tracking general exercise behavior in the adult consumer population. The FBC2 and AW should not be used interchangeably with laboratory gold standards in a research capacity.

TABLE OF CONTENTS

Introduction

The *Physical Activity Guidelines for Americans* recommends adults participate in 150 minutes per week of moderate to vigorous physical activity (MVPA) [1]. However, more than half of US adults do not meet these guidelines [2]. Physical inactivity is a modifiable factor with implications to overall health status of an individual [3] and, by reducing sedentary time, individuals will see benefits in cardiovascular health and improvements in overweight or obesity status [4]. Consumer monitors have caught the eye of researchers by aiding in physical activity goal reinforcement [5]. Activity monitors and their associated applications or websites can aid in goal-setting, goal reflection, and information about the health benefits of physical activity[3]. Further, physical activity monitors can be a tool for promoting physical activity with an individualized approach [6], as this feedback is critical in examination of behavior change with physical activity [7].

Most devices are able to provide information on distance traveled, step count, sedentary time, intensity of activity, energy expenditure (EE), heart rate (HR), and sleep tracking [8]. HR monitoring is a recent feature. The fields of public health, fitness, nutrition, rehabilitation [7], and aging have all sought out to employ consumer devices [8]. It is critical that physical activity monitors are valid in their measurements for their use in aiding weight loss, self-monitoring of both sedentary and active behavior [4], and energy balance interventions [7]. As manufacturers refine their instruments and algorithms utilized, further research is needed to access the accuracy of newly available devices [4]. A few studies have investigated the validity of consumer monitors estimating steps, EE, and active minutes, but there is little evidence existing for validity of HR estimation in the newer models of consumer monitors. [4, 5]. Traditionally, HR monitors have been placed on the chest. Recently, consumer monitors adopted photoplethysmography technology to estimate HR from the wrist using flashing LED light. Currently, little research is available for the change in sensitivity threshold of HR devices, that may be seen with placements other than the chest, such as the wrist [9]. Literature to date has shown greater accuracy in resting heart rate (RHR) measurements over MVPA minutes with new consumer monitors [10]. Additionally, both RHR and MVPA minutes are risk factors for assessing overall cardiovascular health. Therefore, work examining the validity of HR measurements with increasing MVPA minutes, will further the literature in utilizing activity monitors as tools in monitoring and facilitating exercise behavior changes.

Responding to this gap in the literature, the purpose of this study is to examine the validity of HR measurement for the Fitbit Charge 2 (FBC2) and Apple iWatch (AW) with data collected by ActivPAL GT3X+ from the Polar H7 Heart Rate Chest Strap monitor in a short-term free-living setting. This work will fill gaps in the literature on validity of newer models of physical activity monitors measuring HR.

Literature Review

I. Previous Research on Validation of Consumer PA Monitors

There is a wide variety of physical activity monitors commercially available, all focusing on different metrics of activity such as step count, HR monitoring, EE, and activity recognition. When examining step count validity, several studies monitored in laboratory conditions [11-16], while some monitored under free-living conditions [4, 8, 17, 18]. For the validity criterion selection, researchers utilized manual step count [12-15], pedometer [11, 15, 18], Actigraph GT3X [4, 17, 18], or Opal sensors [16]. The monitors examined more frequently in research are the Fitbit Flex, Fitbit One, Fitbit Zip, Jawbone UP, and Jawbone UP24. Fitbit Flex was found to be valid in treadmill activity when compared to manual step counts [11-13, 15] as well as ambulatory household activities [15]. However, the Fitbit Flex was found to have lower accuracy as treadmill speed increased [11], higher error when cycling [15], as well as having difficulties in tracking activity on descending stairs [8]. One study found that hip-based Fitbit One was highly correlated (r=0.97-0.99) to observed step counts across all phases of walking (1.9, 3.0, 4.0mph) and jogging (5.2mph)[13], while another study found the monitor underestimated steps in all of the observed walking speeds[16]. Additionally, Fitbit One was found to significantly underestimate steps during household activities and cycling [15]. Fitbit Zip was found to have a correlation of 0.8 to a pedometer throughout a 24-hour free-living setting [11], despite having a significant underestimation in household activities [15]. The majority of a free-living day is spent sedentary or engaging in light physical activity, therefore monitors, such as Fitbit Zip, that perform best in lab-controlled slow or normal walking activities, were seen to have better accuracy in 24-hour free-living settings [11]. The Jawbone UP was found to have no substantial error on measuring walking/running activity performed on treadmill in Chen et al. study [12].

While, another study found larger error in slower treadmill speeds when examining the Jawbone UP [16]. On the other hand, a study conducted in a 24-hour free-living setting reported that the Jawbone UP24 was correlated at r=0.6 to the criterion (e.g., pedometer) [11], indicating it may not be the best choice for long-term behavior change studies.

In terms of evaluating EE, the majority of researchers chose lab based conditions[13, 15, 19-21], perhaps due to the lack of availability of gold criterions in the field, and some with simulated free-living conditions in lab settings [22-26]. Across the literature, a general finding is that consumer monitors considerably underestimated EE. This finding was consistently observed in studies with Fitbit One [15], Fitbit Zip [15], Jawbone UP24 [15], Fitbit Classic [25], Apple Watch [20, 21], and multiple generations of Mio Alpha [20, 21]. Further, the general consensus is that consumers and researchers need to be weary when utilizing commercially available devices for EE measurements. HR was a relatively trendy feature added to the current line of consumer monitors. It would be assumed that with HR data collected, the accuracy of estimating EE would be higher. There is little known about the EE accuracy from consumer monitors with HR estimation under free-living conditions. Recently, in controlled laboratory conditions, Reddy et al. found the lowest mean EE for FBC2 were observed during activities of daily living (−8.8% [SD 29.2]), and the highest mean error observed during MAX-C (−39.1% [SD 30.6]) and HIIT-C (−41.9% [SD 31.3]) [27]. Garmin (MPE=22.8%) monitor performed better than the FBC2 (MPE=42.7%) during cycle ergometer testing [27]. Yang et al. examined sedentary activity, aerobic exercise, and light intensity physical activity, finding the Fitbit Charge HR to have a mean absolute percent error more than twice the error of the Apple Watch 1 at 32.9% and 15.2%, respectfully [28].

A small portion of the literature on activity monitors validation to date has focused on intensity classification of the devices. Gomersall et al. study examined popular consumer devices hip-based Fitbit One and wrist-based Jawbone UP against validity criterion Actigraph GT3X accelerometry, looking only at MVPA to sedentary behavior [4]. The correlation for steps and MVPA, respectfully, was higher for Fitbit One ($r=.85$, $\rho=.80$) than Jawbone UP ($r=.75$, $\rho=.75$), however, both devices showed systematic bias in mean differences with differences increasing with increasing steps per day[4]. Furthermore, a study examined McRoberts Movemonitor, Jawbone UP, Fitbit One, ActivPAL, Nike+Fuelband, Tractivity, and Sensewear Armband against validity criterion OPAL sensors examining the activities of walking, running, stair climbing and descending, and postural transitions [16]. This study found that the Movemonitor had the lowest error of all devices at all speeds, however, had difficulty discriminating between sitting and standing positioning [16]. A physical activity monitor with high speed detection validity, and yet has difficulty discriminating between daily activities, indicates possible limitations in daily usage by consumers. Compared to step counts and EE, there is currently a gap in the literature of monitors examining activity classification.

II. Previous Research on Heart Rate Estimation in Consumer PA Monitors

When examining HR monitors, researchers chose lab conditions [20, 21, 29-31], yet there does not appear to be any research examining the validity of HR monitors in a free-living setting. The majority of studies utilized electrocardiograms (ECG) as their validation criterion [20, 21, 29, 30, 32]. The utilization of ECG as a criterion is difficult regarding portability and usage in a free-living environment. More work needs to be completed on research grade heart monitoring devices with higher portability and use in free-living research. One solution was with studies such as Barbosa et al. and Engstrom et al., who tested chest strap monitors Polar RS80063 and

Polar RS400, respectively [29, 30] and found the research grade chest strap monitors to be valid compared to ECG [29, 30]. Thus, the validity in HR measurements have made chest straps popular in those interested in self-monitoring HR from an adult consumer population.

As HR monitors become more commercially available, research on wrist-strap monitors is increasing. As well, imputing HR measurements into EE equations of the devices is prevalent [20]. Research, thus far, has focused on validity of popular consumer monitors during participation in a variety of common physical activity modes including treadmill walking [20, 27, 33, 34], treadmill running [20, 27, 34], elliptical [33], stationary bike [20, 33], high intensity interval training [27], and activities of daily living [27]. The Apple iWatch was found to have high HR measurement validity under controlled laboratory conditions during cycling [20, 33, 35], no statistical difference from ECG during treadmill walking and running ($p=0.22$) [33]. In addition, the Apple iWatch was found to have a mean absolute percent difference to Polar H7 HR strap between 1.14% and 6.70% across all treadmill protocol speeds [34], and similar agreement during vigorous intensity activity as other physical activity intensities [33]. Data on Fitbit devices varied across the literature. FBC2 had acceptable HR accuracy during low intensity exercise [27] but was found to underestimate HR measurements at higher intensity physical activity regimens [35]. The Dooley et al. study found during treadmill stages, the Fitbit Charge Heart Rate had significantly lower HR measurements during baseline ($p<0.001$, $d=0.15$), vigorous intensity ($p<0.001$, $d=0.31$), and recovery ($p<0.001$, $d=0.13$), while significantly higher HR at light intensity activity ($p<0.001$, $d=0.68$) [34]. Fitbit Blaze also underestimated HR, with high error during elliptical with arm movement [33].

Some studies found a systematic bias across all wearable devices. As lab-controlled protocol intensity increased, there was a higher overall error in HR measurements [20, 27, 35], while one study found through Bland-Altman analysis that variability in HR was not influenced by HR magnitude [33].

Further, as more advanced technology is incorporated into physical activity monitors and as they become commercially available and affordable, some researchers highlight concern that skin tone may interfere with the photoplethysmography technology for HR detection. One study found that the covariate of darker skin tone positively correlated with increased error of HR detection among the tested devices [20]. In agreement with this finding, another study proposed that participants with less photosensitive skin, had increased problems with device functionality [36].

III. Laboratory (and Simulated Free-Living) vs. Free-living Setting

Consumer activity monitor validation studies are often conducted in two settings: laboratory and free-living. A well-controlled laboratory setting eliminates the random errors, providing a channel to study the monitor’s validity for certain types of activity. However, controlled laboratory settings may have limited real-life applicability [4, 37]. Some studies instead examined devices from both a laboratory-based and free-living setting to more fully understand the capabilities of certain activity trackers [37]. Activity monitors are designed to track daily free-activity for general consumers. Free-living setting validity allows for further intervention research to be conducted using consumer monitors with known validity [4]. However, equations developed under laboratory conditions may not transfer to free-living environments where the monitors are utilized by consumers [22]. The current consumer monitor validation literature consists mostly of studies designed in laboratory settings and only a few were conducted in free-living setting.

The current free-living setting validation studies primarily focused on step counts [11, 38]. Yet, there are some limitations with using commercially available devices in free-living research studies. Long check in points result in increased chance of misplacement of the device [39] or battery life issues [17], as well as problems with synchronization of the device to record the data, resulting in inaccurate recounts of activity [39]. As such, some studies chose to examine devices from both a laboratory-based and free-living setting to more fully understand the capabilities of certain activity trackers for future use in research [37].

IV. MVPA and HR Measurements

Preliminary findings of recent photoplethysmography technology consumer monitors measuring HR show greater accuracy in readings of RHR compared to HR measurements during MVPA minutes [10]. Degroote et al. study examined accuracy in MVPA measurements in a two-day free-living setting examining overall accuracy by day, as well as on a 15-minute level. This allowed investigation of the potential of the devices to correctly situate physically active behavior over time and to provide exact real time feedback on physical activity behavior [40]. The Fitbit Charge was found to underestimate MVPA time on the day level by 30%, and overestimate MVPA on the 15-minute level by 20% [40]. Further, an examination of other Fitbit models by Reid et al. study found the Fitbit Flex and Fitbit One, compared to Actigraph GT3X+ accelerometer data, overestimated the time spent in MVPA [41]. Additionally, there is a concern that consumer physical activity monitors are utilizing different MVPA level cutoffs than current gold standards used in accelerometers, making it difficult to compare the MVPA measurements [40].

Further research is necessary to examine validity of HR measurements across all intensity and modes of physical activity in free-living settings. The primary purpose of this study was to address the gap in the literature in validity of new wrist-worn monitors, FBC2 and AW, in HR measurements against criterion Polar H7 Bluetooth Heart Rate Chest Strap monitor over a 24-hour free-living setting.

Methods

Participants

Healthy adults aged 18-55 years participated in the study, where 16 (8 male, 8 female) successfully obtained minute level HR data for FBC2 and 33 (12 male, 21 female) successfully obtained HR data from the AW. Participants were recruited through fliers, emails, and classroom announcements throughout the University of Vermont and UVM Medical Center resulting in the majority of participation being composed of college-aged adults (18-22 years old). Each participant completed a phone screening of the Physical Activity Readiness Questionnaire (PAR-Q) with additional questions. Exclusion criteria included any known metal allergy, tattoos on either wrist, any current mobility assistance, and a "yes" response on the PAR-Q. Upon their first visit, each participant signed informed consent before beginning the protocol. The study was approved by the institutional review board at the University of Vermont.

Instruments

Criterion Measure:

The Polar H7 Bluetooth Heart Rate Chest Strap monitor was utilized to obtain criterion HR measures. The monitor was strapped to the participant's chest with the electrode area positioned just below pectoral muscles and placed firmly against the skin. The device sent data through Bluetooth to the Actigraph GT3X+ worn on the wrist of the participant. This device is commonly utilized in consumer monitor research to quantify physical activity in free-living settings [28].

Consumer Activity Monitors:

The present study examined the HR data collected from FBC2 (2017 version, Fitbit Inc., San Francisco, CA) and AW (Apple Inc., Cupertino, CA). Both activity monitors track steps, HR, active minutes, and EE. Additionally, both FBC2 and AW allow one to select the mode of physical activity, track sleep stages, and lead guided meditation based on detected HR. In terms of HR capture, FBC2 uses photoplethysmography to automatically and continuously capture real-time minute-by-minute HR data. FBC2 uses "PurePulse" technology which, using photodiodes, allows for light absorption to be utilized to measure HR. The green LED light sensor reflects off the skin of the wrist to detect that change in blood flow to the capillaries with each heart beat and extrapolates given beats per minute. AW utilizes an ultrathin silicon carbon nitride layer applied to sapphire crystals to read electrical impulses in ones' radial artery.

Protocol

Data collection began March 2017 and ended November 2017. Participants were required to attend two visits at the Rowell laboratory. Anthropometric data was collected at visit one. Height was measured using a stadiometer. Blood pressure and RHR were recorded three times using an Omron 10 Series Wireless Bluetooth Upper Arm Blood Pressure Monitor. Body composition and weight were obtained through a bioelectrical impedance analysis unit (BIA), the SC-331S Total Body Composition Analyzer with Column (2016 Model, Tanita Corporation of America Inc., Arlington Heights, IL). The TANITA Body Composition Analyzer is indicated for use in the measurement of weight and impedance, and the estimation of body mass index (BMI), total body fat percent, total body water percent and weight, muscle mass (skeletal and smooth), bone mass, visceral fat rating with healthy range, basal metabolic rate (BMR), physique rating, metabolic age, and target body fat percent with predicted weight and fat mass.

Physical activity monitors FBC2, Fitbit Alta, and AW were initialized with participant anthropometric and demographic data. Participants were instructed about the location of the seven fitness monitors: AW was distal to wrist Actigraph and were positioned on the left wrist; FBC2 was distal to Fitbit Alta on right wrist; a Polar H7 Bluetooth Heart Rate Chest Strap was worn firmly on chest; and a waist Actigraph and DIGI-walker CW2000 pedometer was hooked onto clothing. Instructions on the requirements of the 24-hour free living measurement period outlined the need to wear the FBC2, AW, and Fitbit Alta at or before midnight of the selected protocol day. The additional monitors were put on once out of bed for the day. The participants filled out time logs for when all monitors were on or off their body for the 24-hour measurement period including for showers, water activities, or sleeping bouts. Participants wore the monitors from midnight to midnight of the following day and were instructed to go about their normal activities of daily living. Participants returned the day following the monitored period for laboratory visit two. All fitness monitors were returned as well as the time log of when monitors were worn. The AW, FBC2, and Fitbit Alta were synced to the protocol phone to make sure the data was captured from the monitor period. Participants filled out a self-guided physical activity recall (PAR24) using the ACT24 respondent site.

Data Acquisition and Processing

The criterion HR data from the Polar H7 Bluetooth Heart Rate Chest Strap were downloaded in minute-by-minute format using Actigraph software. For the FBC2, minute-by-minute HR data was obtained through a third-party website Fitabase (Small Steps Labs LLC., San Diego, CA). Two sets of Fitbit monitors were utilized during the study. One monitor failed to sync with Fitabase, accounting for our sample size difference between FBC2 and AW data collection. AW data is not collected in a fixed sampling frequency and the nature of the study (free-living) did not allow for us to manually collect HR data without interruption to the study protocol. Thus, the time collected was matched to the criterion time to carry out the HR validation. HR data from criterion were merged with data collected from the two consumer monitors (e.g., FBC2 and AW series 2) at the minute level. The data set analyzed removed any minute count where the criterion Polar HR Chest Strap monitor recorded either a zero with activity recorded, or a high HR, with no activity recorded as the criterion malfunctioned in capturing the data. All data was processed following the 24-hour free-living monitoring period.

Statistical Analyses

Participant demographic and anthropometric data were summarized using descriptive statistics. Correlation analysis utilizing Pearson's Correlation Coefficient was completed to examine the relationship between the criterion and the consumer activity monitors minute level agreements in measurements of HR. This relationship was examined at three intensity levels of physical activity: sedentary behavior (SB), light physical activity (LPA), and moderate-to-vigorous physical activity (MVPA). We applied the Freedson 1998 cut off criterion using waistline Actigraph data to classify the intensity with counts less than 100 counts/min as sedentary activity, any activity between 100-1951counts/min as LPA, and intensity is above 1951 counts/min as MVPA [42]. Minute by minute measurements were evaluated using mean percent errors (MPE), mean absolute percent errors (MAPE), and root-mean-square errors (RMSE). MPE was calculated through averaging the individual minute-by-minute error. MAPE was calculated through absolute percent error averages. RMSE was calculated through the square root of the mean square error. Bland-Altman plots were used to evaluate the mean difference between the criterion and FBC2 and AW monitors at three intensity levels of physical activity.

Results

The study included a total of fifty-two participants. Participants' demographic and anthropometric characteristics are provided in Appendix Table 1. Participants were classified as healthy adults, with a normal RHR range classified as 68-72 bpm and normal systolic blood pressure of less than 120 mmHg, and diastolic blood pressure of less than 80 mmHg [1]. However, complete data was not obtained from all individuals due to technical issues, or a participant's inability to wear devices for the 24-hour period because of interference with normal daily activities. A total of 16 individuals completed FBC2 data from 24-hours of free-living resulting in 14,988 time points to be analyzed [mean RHR of 68±10.5 bpm, SBP of 110.9±11.4 mmHg and DBP 66.5±9.9 mmHg]. The relative percentages of time spent in SB, LPA, and MVPA were 52.1%, 43.76%, and 4.12%, respectively for the FBC2. A total of 33 individuals completed AW data from 24 hours of free-living resulting in 5,109 time points to be analyzed [mean RHR of 71.7±10.5 bpm, SBP of 112.0± 11.9 mmHg, and DBP of 69.4± 9.4 mmHg]. The relative percentages of time spent in SB, LPA, and MVPA were 47.84%, 40.61%, and 11.55%, respectively for the AW. [1]

A summary of HR validity indicators from the FBC2 are displayed in Table 2. The FBC2 overestimated heart rate with a MPE of 2.7%, 4.9% and 13% for sedentary, moderate, and vigorous levels of physical activity, respectively. There was a positive direct relationship between physical activity, SB, LPA, MVPA, and the error as shown through an increase in MAPE values of FBC2 4%, 10%, 14% and AW 7%, 10%, 16%. The MAPE were smaller in magnitude for all modes of physical activity for the FBC2 compared to the AW. In examining Pearson Correlation values, both monitors had a strong relationship during sedentary behavior (r_{FBC2} =0.90, $p<0.0001$ and r_{AW}=0.73, $p<0.0001$). Regarding light physical activity, FBC2 had a strong relationship while AW had a moderate strength correlation (r_{FBC2}=0.70, $p<0.0001$ and

r_{AW}=0.56, p<0.0001). Vigorous physical activity had a moderate relationship for both FBC2 and AW (r_{FBC2}=0.49, p <0.0001 and r_{AW}=0.49, p<0.0001).

Bland-Altman plot analyses display the distribution of error and aid in identifying systemic bias of devices. The Bland-Altman plot findings are presented in Appendix Figure 1 and Appendix Figure 2 for FBC2 and AW, respectfully. Overall, both FBC2 and AW tended to underestimate HR values. The plot spreads in Figure 1 and Figure 2 show a direct relationship between variability increasing as HR increases, due to increase in physical activity intensity. The plots revealed the narrowest 95% limits of agreement for the rest mode of FBC2 (difference=19.5 bpm) and the widest limits of agreement for FBC2 during MVPA (difference=71.1 bpm) and AW during MVPA (difference=78.6 bpm). There is bias seen in Figure 1D and Figure 2D as the average difference is non-zero (13.9 bpm, 15.5 bpm). For the estimation of HR, analysis revealed a systematic bias with mean difference for FBC2 in variation being dependent on magnitude of reported HR, with difference value spreading as HR increases (Figure 2A).

Discussion

This study was conducted to broaden the literature examining validity of newly released consumer physical activity models in a healthy adult population. In whole, FBC2 and AW were found to have reasonable validity of HR measurements in healthy adults. These findings align with studies to date in laboratory settings for both the FBC2 [27] and AW [20, 33-35]. Bland-Altman analysis revealed a net underestimation of HR measurements by both the FBC2 and AW. FBC2's underestimation of HR measurement is consistent with the literature across various physical activity modes in a laboratory setting including treadmill [27], HIIT [27], activities of daily living [27], resistance exercise [35], and stationary cycling [35, 43]. Boudreaux et al. found similar trends in Bland-Altman analyses with increasing mean differences and 95% confidence intervals due to continuous HR underestimation by FBC2 [35]. In contrast, Dooley et al. found underestimation or overestimation of HR measurements compared to criterion Polar Heart Rate Strap monitor to be dependent on intensity of physical activity [34]. Further research is necessary to discern a cause behind these error variations. It is also important to note that the literature is fairly limited on these specific models as they are newer devices.

HR Accuracy during moderate to vigorous intensity physical activity

Our study focused on validity of the physical activity monitors across sedentary behavior, LPA, and MVPA. Both the FBC2 and AW HR measurements were found to be valid when engaging in sedentary behavior. High validity when engaging in light physical activity levels is consistent with literature to date for the FBC2 [35] and AW [20, 33, 35]. A recent study by Dooley et al. found that AW did not have significant difference from Polar Heart Rate monitor at baseline ($\text{Mean}_{\text{polar}}$= 72.99, Mean_{AW}= 73.07, p=0.76), however had significantly lower HR readings during light activity ($\text{Mean}_{\text{polar}}$= 92.45, Mean_{AW}= 89.19,p=0.03, d=-0.25) and moderate intensity ($\text{Mean}_{\text{polar}}$= 106.84, Mean_{AW}= 101.01, d= −0.35) [34]. This aligns with our findings that higher HR measurements tend to be underestimated by consumer devices when compared to gold standard.

Compared to Polar H7 Heart Rate Strap criterion measure, FBC2 was found to have slightly higher validity than AW for HR measurements when engaging in MVPA levels. A recent study by Boudreaux et al. found that AW series 2 had higher HR measurement validity compared to Polar HR strap monitor using a cycle ergometer (R=0.80-0.90) and during resistance exercise ($R = 0.72$) [35]. Reddy et al. also found differences in accuracy of varying activities for treadmill maximum testing (R=0.94) versus HIIT on a cycle ergometer (0.46) [27]. Our study did not specifically focus on mode of exercise engaged in during the 24-hour free living period, however this suggests that different modes of physical activity may impact the accuracy of reported HR measurements. Examination of outlier HR measurement data from the criterion device compared to 24-hour activity recall logs completed by participants showed the errors mainly fell at the initiation or termination of MVPA or during exercise with rapid arm movements such as basketball or racquetball.

Overall, both FBC2 and AW displayed a moderate relationship with data classified as MVPA. The decline in accuracy with increasing intensity of physical activity in devices has been noted in the literature for FBC2 [27, 35] and AW [35]. In contrast, the Gillinov et al. study found that there was no impact on accuracy of HR measurements with an increase in magnitude of HR that accompanies higher intensity activities [33]. One suggested explanation for an impact on HR measurement accuracy as intensity levels increase is with sustained movement, the contact between the photoplethysmography sensor and the skin lessens, therefore leading to the disruption of signaling and decreasing quality of the data [27]. The HR measurement may also be impacted by how much the arm with the physical activity monitor is involved in the exercise. For example, running is a very uniform motion while weight lifting activities includes more jerking and twisting motions, potentially affecting the ability of the sensor to capture data. This decline in accuracy is seen in previous Fitbit models including Fitbit One and Fitbit Flex which were found to overestimate MVPA minutes [41]. These monitors may have utilized a different biofeedback model accounting for an overestimation of MVPA minutes differing from our findings on FBC2.

Photoplethysmography in Consumer Physical Activity Devices

There are a variety of concerns with utilizing the photoplethysmography technology in the devices in that the technology has greater error in certain subsets of the general population. Covariates including skin pigmentation, larger waist circumference, and larger BMI were found to positively correlate with an increase in HR measurement error rates across multiple devices [20]. Our study found no significance in BMI and HR measurement error for the FBC2 (r=-0.02429) and the AW (r=0.02382). However, our study focused on adults classified as healthy and further research is necessary to determine if these covariates effect HR measurement error in the FBC2 and AW. These covariates are particularly concerning as many individuals may be looking to utilize these devices to help reach weight loss goals. Additionally, the Shcherbina et al. study found that males had 4% higher error in their HR readings across all devices and tasks [20]. Our study excluded individuals with wrist tattoos as they have been found to interfere with device readings.

Photoplethysmography relies on detection of blood flow to the limb the device is worn on, so for participants that have decreased blood flow to upper extremities, the devices may not be accurate. This is one major drawback to this technology as populations looking to increase physical activity to combat cardiovascular disease may have difficulties getting accurate feedback from the monitors. HR measurements being lower than they should leads to concerns regarding frustration and discontinuation of physical activity behavior change. The monitor underestimating HR measurements is an issue with high risk populations who then work harder to get their HR up when it is already high and may reach dangerous levels for their current health status. In addition, for wrist-worn monitors, the tightness of the band influences the accuracy of the HR measurements. In our study, we instructed participants to wear the devices to a tightness level that felt comfortable, not too restricting and not too loose on the wrist.

Strengths

To our knowledge, this is one of the first studies to examine HR validation for FBC2 and AW in a short-term, free-living setting. Our short-term, free-living study was the next step in examining validity of consumer devices beyond a controlled laboratory setting. The strengths of our study include the free-living procedure as it is more in line with the purpose of having reliable physical activity monitors available to the general public. Consumer physical activity monitors need to be accurate for exercise bouts as well as activities of daily living.

Limitations

As this study's participants were classified as healthy adults, further research is needed to assess validity of the FBC2 and AW in other populations that may be looking to utilize the device to monitor activity. Populations that may have variation in device accuracy that could be focused on include: underweight or overweight populations, youth populations, and geriatric populations. One limitation of this study is the sample size difference between FBC2 and AW data collection due to failure of one FBC2 monitor to connect to Fitabase for data analysis. It is also important to note that only 4.12% for the FBC2 and 11.55% for the AW of data was spent in MVPA. As there was less data to analyze, the underestimation trend may be due to a smaller sample size compared to that of SB and LPA in this study. Additionally, a free-living setting is not as thoroughly controlled as a laboratory-based setting. For example, participants were unmonitored during the 24-hour period, therefore accurate device placement cannot be guaranteed. There were instances of participant data being incomplete due to interference of wearing the device with their daily lives, for example, working as a dishwasher at a restaurant. This interference with daily lives may be necessary to include as an inclusion criterion in future studies to obtain complete datasets. Due to the free-living setting, we cannot guarantee that participants wore the monitors in the correct manner. For example, if the band was not tight enough on the wrist, the photo-reliant technology may not be reading as accurately as it should, especially if the activity being performed involves prolonged wrist movements.

Future Research

The trend of physical activity wearable devices to be worn on the wrist brings about the concern of potential spurious results from upper extremity movement. Our study did not focus on the potential of activities involving prolonged wrist movements impacting the accuracy of wrist-worn devices, however, research on newer devices has found evidence of differences in HR measurements based on which wrist it was worn [27]. Further research is necessary to investigate spurious measurements and whether selecting the exercise mode with the device lessens the measurement error from laboratory gold standards.

As more research is completed, a net underestimation of these devices at higher levels of intensity may be found and can be used to educate consumers on the amount of underestimation they may experience at higher levels of PA intensity. Although our study found a net underestimation at higher levels of activity, the majority of daily time is spent in SB or LPA levels, of which both the FBC2 and AW were valid. Therefore, the devices can provide useful estimation for everyday use by healthy adults.

Current physical activity monitors allow for anthropometric and demographic data to be entered in the device to be factored into physical activity recommendations provided by the devices and corresponding application or website. Our study entered in weight, height, gender, and age on the Fitbit application and Apple device application. Future research should examine whether additional anthropometric measurements reduce the error in these devices [34].

The knowledge of which monitor may work best for a population or setting will be beneficial for consumers and professionals aiding in exercise behavior determining which device will work best for them in their respective physical activity goals.

Conclusion

The FBC2 and AW should be used with caution, especially with higher levels of exercise, however they have reasonable validity for tracking general exercise behavior changes in consumer population. The overall cost, comfort, style, and additional features of the FBC2 and AW make them best suited for the consumer population looking to have greater awareness of their health behaviors. The FBC2 and AW should not be used interchangeably with laboratory gold standards in a research capacity.

Consumer physical activity monitors now boast a wide range of health behavior settings, including tracking for sleep, a “relax” personalized meditation function, and beyond. Once the validity of physical activity monitors is assessed, the validity of HR measurements extending to these functions should be evaluated, especially considering the feedback may be reliant on the HR measurement quality of the devices. Validation of tracking various health behaviors would be beneficial to the consumer population in having the means to track and set goals for additional aspects of their health.

As companies releasing new consumer activity tracking devices do not generally release the formulas for calculations on steps, HR, and EE, among others, researchers will need to continue to evaluate the efficacy of current devices in their ability to provide accurate information to consumers.

Acknowledgements

I would like to express my thanks to my advisor, Dr. Yang Bai, for her support and guidance throughout this entire process. I am grateful to have been mentored by someone so patient and encouraging. Additionally, I would like to thank my committee members Dr. Susan Kasser and Dr. Connie Tompkins for taking their time to add their views into the discussion. Finally, I would like to acknowledge my peers and family members, as without their unwavering support, completing this project would not have been possible.

References

1. Piercy, K.L. and R.P. Troiano, *Physical Activity Guidelines for Americans From the US Department of Health and Human Services.* Circ Cardiovasc Qual Outcomes, 2018. **11**(11): p. e005263.
2. King, A.C. and J.F. Sallis, *Why and how to improve physical activity promotion: lessons from behavioral science and related fields.* Prev Med, 2009. **49**(4): p. 286-8.
3. Vandelanotte, C., et al., *What kinds of website and mobile phone-delivered physical activity and nutrition interventions do middle-aged men want?* J Health Commun, 2013. **18**(9): p. 1070-83.
4. Gomersall, S.R., et al., *Estimating Physical Activity and Sedentary Behavior in a Free-Living Context: A Pragmatic Comparison of Consumer-Based Activity Trackers and ActiGraph Accelerometry.* J Med Internet Res, 2016. **18**(9): p. e239.
5. *<Endeavour-Partners-Wearables-White-Paper-how behavior change science offers the secret to long-term engagement.pdf>.*
6. Conroy, D.E., C.H. Yang, and J.P. Maher, *Behavior change techniques in top-ranked mobile apps for physical activity.* Am J Prev Med, 2014. **46**(6): p. 649-52.
7. Lyons, E.J., et al., *Behavior change techniques implemented in electronic lifestyle activity monitors: a systematic content analysis.* J Med Internet Res, 2014. **16**(8): p. e192.
8. Kaewkannate, K. and S. Kim, *A comparison of wearable fitness devices.* BMC Public Health, 2016. **16**: p. 433.
9. Tudor-Locke, C., et al., *Comparison of pedometer and accelerometer measures of free-living physical activity.* Med Sci Sports Exerc, 2002. **34**.
10. Gorny, A.W., et al., *Fitbit Charge HR Wireless Heart Rate Monitor: Validation Study Conducted Under Free-Living Conditions.* JMIR Mhealth Uhealth, 2017. **5**(10): p. e157.
11. An, H.S., et al., *How valid are wearable physical activity trackers for measuring steps?* Eur J Sport Sci, 2017. **17**(3): p. 360-368.
12. Chen, M.D., et al., *Accuracy of Wristband Activity Monitors during Ambulation and Activities.* Med Sci Sports Exerc, 2016. **48**(10): p. 1942-9.
13. Diaz, K.M., et al., *Fitbit(R): An accurate and reliable device for wireless physical activity tracking.* Int J Cardiol, 2015. **185**: p. 138-40.
14. Hickey, A., et al., *Validity of Activity Monitor Step Detection Is Related to Movement Patterns.* J Phys Act Health, 2016. **13**(2): p. 145-53.
15. Nelson, M.B., et al., *Validity of Consumer-Based Physical Activity Monitors for Specific Activity Types.* Med Sci Sports Exerc, 2016. **48**(8): p. 1619-28.
16. Storm, F.A., B.W. Heller, and C. Mazza, *Step detection and activity recognition accuracy of seven physical activity monitors.* PLoS One, 2015. **10**(3): p. e0118723.
17. Schneider, M. and L. Chau, *Validation of the Fitbit Zip for monitoring physical activity among free-living adolescents.* BMC Res Notes, 2016. **9**(1): p. 448.
18. Tully, M.A., et al., *The validation of Fitbit Zip™ physical activity monitor as a measure of free-living physical activity.* BMC Research Notes, 2014. **7**(1): p. 952.
19. *<Montgomery 2009-JSCR-Validation of HR monitor-based predictions of oxygen uptake and energy expenditure.pdf>.*
20. Shcherbina, A., et al., *Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort.* 2016.
21. Wallen, M.P., et al., *Accuracy of Heart Rate Watches: Implications for Weight Management.* PLoS One, 2016. **11**(5): p. e0154420.
22. WELK, G.J., S. N. BLAIR, K. WOOD, S. JONES, and R. W. THOMPSON, *<Welk_2000_MSSE_A comparative evaluation of three accelerometers.pdf>.* Journal of the American College of Sports Medicine, 2000: p. p. 489-497.

23. Dannecker, K.L., et al., *A comparison of energy expenditure estimation of several physical activity monitors.* Med Sci Sports Exerc, 2013. **45**(11): p. 2105-12.
24. Lee, J.M., Y. Kim, and G.J. Welk, *Validity of consumer-based physical activity monitors.* Med Sci Sports Exerc, 2014. **46**(9): p. 1840-8.
25. Sasaki, J.E., et al., *Validation of the Fitbit wireless activity tracker for prediction of energy expenditure.* J Phys Act Health, 2015. **12**(2): p. 149-54.
26. Tucker, W.J., et al., *Validity and reliability of Nike + Fuelband for estimating physical activity energy expenditure.* BMC Sports Sci Med Rehabil, 2015. **7**: p. 14.
27. Reddy, R.K., et al., *Accuracy of Wrist-Worn Activity Monitors During Common Daily Physical Activities and Types of Structured Exercise: Evaluation Study.* JMIR Mhealth Uhealth, 2018. **6**(12): p. e10338.
28. Bai, Y., et al., *Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR.* J Sports Sci, 2018. **36**(15): p. 1734-1741.
29. Barbosa, M.P., et al., *Comparison of Polar(R) RS800G3 heart rate monitor with Polar(R) S810i and electrocardiogram to obtain the series of RR intervals and analysis of heart rate variability at rest.* Clin Physiol Funct Imaging, 2016. **36**(2): p. 112-7.
30. Engström, E., et al., *Comparison of heart rate measured by Polar RS400 and ECG, validity and repeatability.* Advances in Physiotherapy, 2012. **14**(3): p. 115-122.
31. Stahl, S.E., et al., *How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough?* BMJ Open Sport Exerc Med, 2016. **2**(1): p. e000106.
32. Parak, J. and I. Korhonen, *Evaluation of wearable consumer heart rate monitors based on photopletysmography.* Conf Proc IEEE Eng Med Biol Soc, 2014. **2014**: p. 3670-3.
33. Gillinov, S., et al., *Variable Accuracy of Wearable Heart Rate Monitors during Aerobic Exercise.* Med Sci Sports Exerc, 2017. **49**(8): p. 1697-1703.
34. Dooley, E.E., N.M. Golaszewski, and J.B. Bartholomew, *Estimating Accuracy at Exercise Intensities: A Comparative Study of Self-Monitoring Heart Rate and Physical Activity Wearable Devices.* JMIR Mhealth Uhealth, 2017. **5**(3): p. e34.
35. Boudreaux, B.D., et al., *Validity of Wearable Activity Monitors during Cycling and Resistance Exercise.* Med Sci Sports Exerc, 2018. **50**(3): p. 624-633.
36. Spierer, D.K., et al., *Validation of photoplethysmography as a method to detect heart rate during rest and exercise.* J Med Eng Technol, 2015. **39**(5): p. 264-71.
37. Kooiman, T.J., et al., *Reliability and validity of ten consumer activity trackers.* BMC Sports Sci Med Rehabil, 2015. **7**: p. 24.
38. Ferguson, T., et al., *The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study.* Int J Behav Nutr Phys Act, 2015. **12**: p. 42.
39. Harrison, D., et al., *Tracking physical activity*, in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct*. 2014. p. 699-702.
40. Degroote, L., et al., *The Accuracy of Smart Devices for Measuring Physical Activity in Daily Life: Validation Study.* JMIR Mhealth Uhealth, 2018. **6**(12): p. e10972.
41. Reid, R.E.R., et al., *Validity and reliability of Fitbit activity monitors compared to ActiGraph GT3X+ with female adults in a free-living environment.* J Sci Med Sport, 2017. **20**(6): p. 578-582.
42. Freedson, P.S., E. Melanson, and J. Sirard, *Calibration of the Computer Science and Applications, Inc. accelerometer.* Med Sci Sports Exerc, 1998. **30**.
43. Benedetto, S., et al., *Assessment of the Fitbit Charge 2 for monitoring heart rate.* PLoS One, 2018. **13**(2): p. e0192691.

Appendix

Table 1. Descriptive demographics of FBC2 and AW participants

	FBC2 (N=16)		AW (N=33)	
	Mean	SD	Mean	SD
Height (cm)	171.1	11.4	169.3	10.8
Weight (kg)	68.4	13.2	66.9	12.1
Muscle Mass (Kg)	56.6	14.9	54.2	14.8
RHR (bpm)	68.1	10.5	71.7	10.5
SYS (mmHg)	110.9	11.4	112.0	11.9
DIA (mmHg)	66.5	9.9	69.4	9.4

*Abbreviations: SD, standard deviation, RHR, resting heart rate, SYS, systolic blood pressure, DIA, diastolic blood pressure.

Table 2. Comparison of HR estimation (bpm) from FBC2 to Criterion

	Polar H7 HR Strap Mean HR Measurements (SD)	FBC2 Mean HR Measurements (SD)	MPE (SD)	MAPE (SD)	RMSE	Correlation
SB (N=7811)	59.63 (10.58)	61.54 (11.53)	2.7% (7.1%)	4% (6.1%)	8.94	0.90**
LPA (N=6559)	70.71 (13.69)	74.77 (12.24)	4.9% (14%)	10% (10%)	11.34	0.70**
MVPA (N=618)	108.90 (17.66)	125.59 (17.99)	13% (13%)	14% (12%)	22.33	0.49**

*Abbreviations: SB, Sedentary Behavior, LPA, light physical activity, MVPA, moderate-vigorous physical activity, MPE, mean percent error, MAPE, mean absolute percent error, RMSE, root mean square error.

**Statistically significant values p<0.001.

Table 3. Comparison of heart rate estimation from AW to Criterion

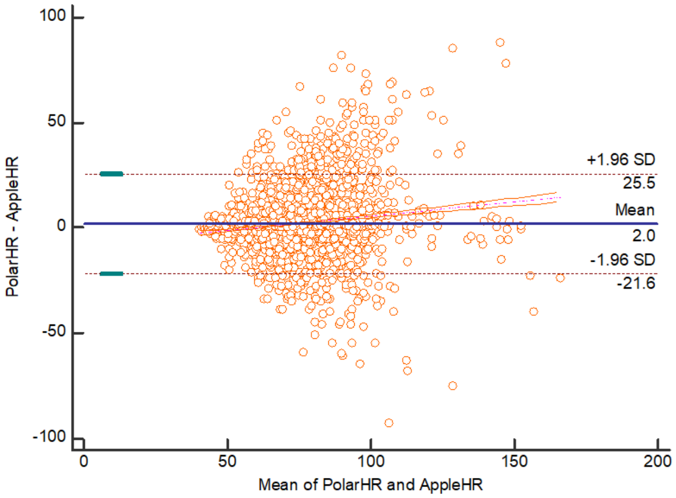
	Polar H7 HR Strap Mean HR Measurements (SD)	AW Mean HR Measurements (SD)	MPE (SD)	MAPE (SD)	RMSE	Correlation
SB (N=2444)	66.59(12.32)	66.74(11.78)	-1.1% (13%)	7% (11%)	5.33	0.73**
LPA (N=2075)	75.63(11.97)	73.55 (11.84)	1.8% (15%)	10% (11%)	10.97	0.56**
MVPA (N=590)	94.34(19.34)	83.94(19.67)	9.6% (19%)	16% (14%)	24.54	0.49**

*Abbreviations: SB, sedentary behavior, LPA, light physical activity, MVPA, moderate-vigorous physical activity, MPE, mean percent error, MAPE, mean absolute percent error, RMSE, root mean square error.

**Statistically significant values p<0.001.

Figure 1. Bland-Altman Plot comparisons of mean difference in HR measurements of Polar H7 Heart Rate Strap monitor and AW

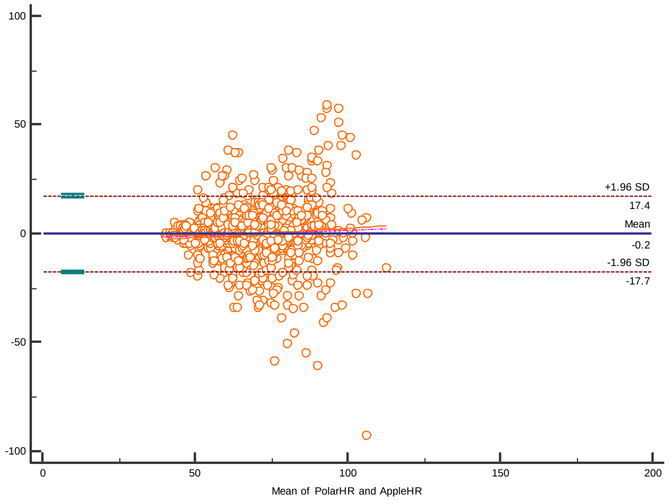
A. All HR measurements for AW



95% CI 1.6429 to 2.3016

Regression Equation y = -7.7870 + 0.1347 x

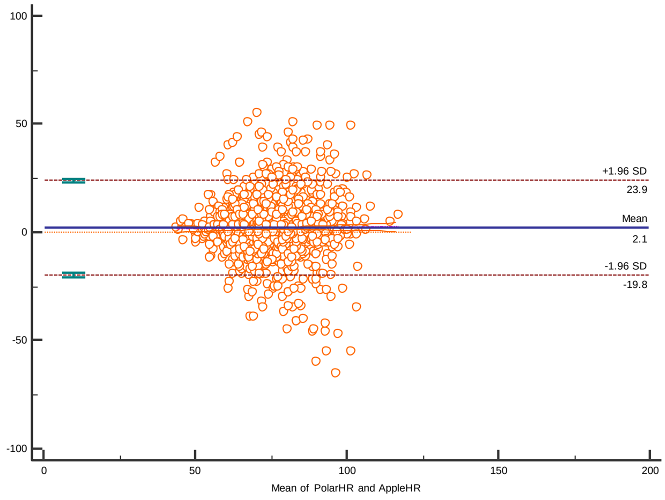
B. At Rest HR Measurements



95% CI -0.5118 to 0.1976

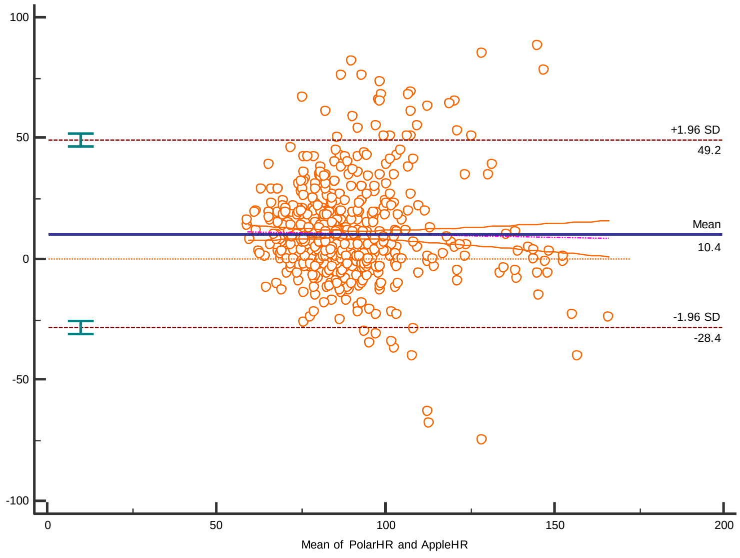
Regression Equation y = -3.6110 + 0.05181 x

C. Light intensity activity HR measurements



95% CI	1.6024 to 2.5624
Regression Equation	y = 1.0343 + 0.01405 x

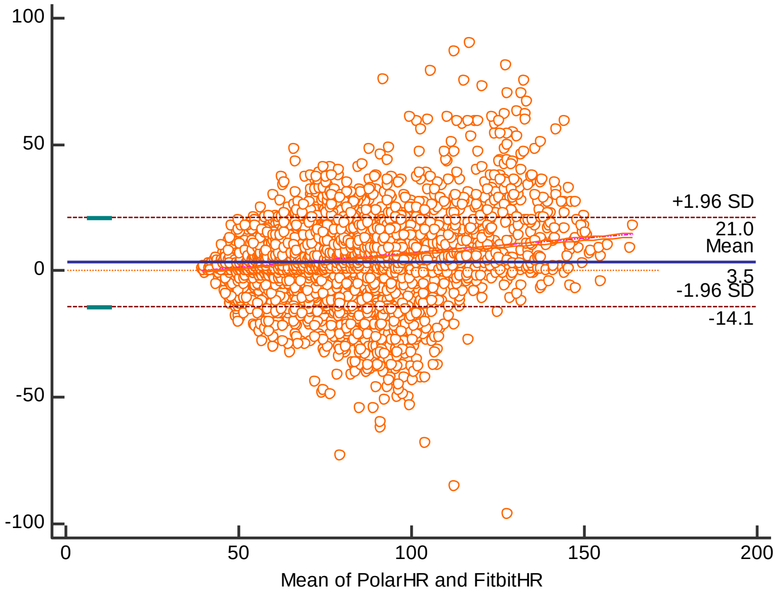
D. Moderate-Vigorous activity HR measurements



95% CI	8.8061 to 12.0041
Regression Equation	y = 12.4224 + -0.02263 x

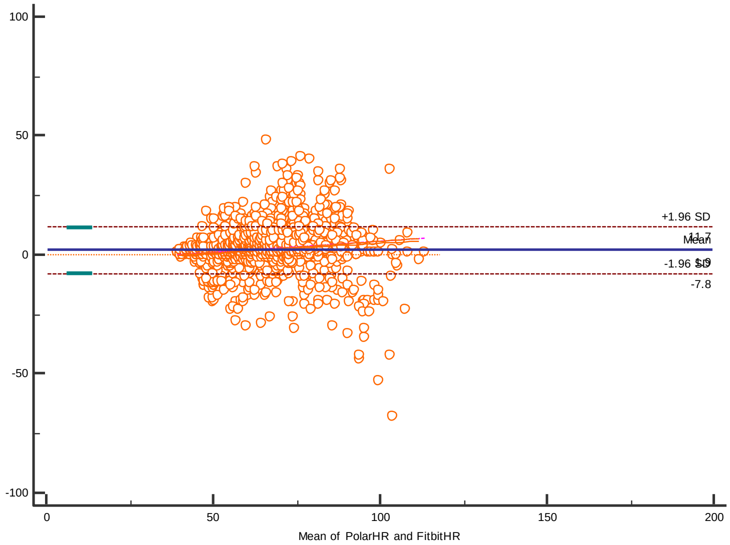
Figure 2. Bland-Altman Plot comparisons of mean difference in HR measurements of Polar H7 Heart Rate Strap monitor and AW

A. All HR measurements for FBC2



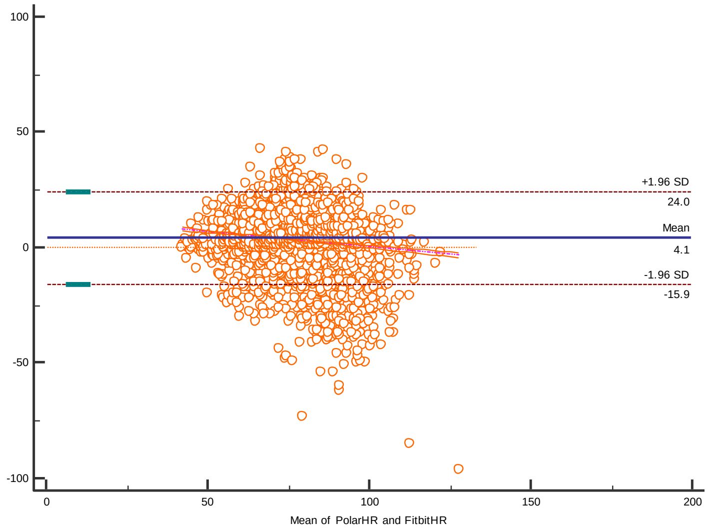
95% CI	3.3199 to 3.6070
Regression Equation	y = -4.4686 + 0.1162 x

B. At rest HR measurements



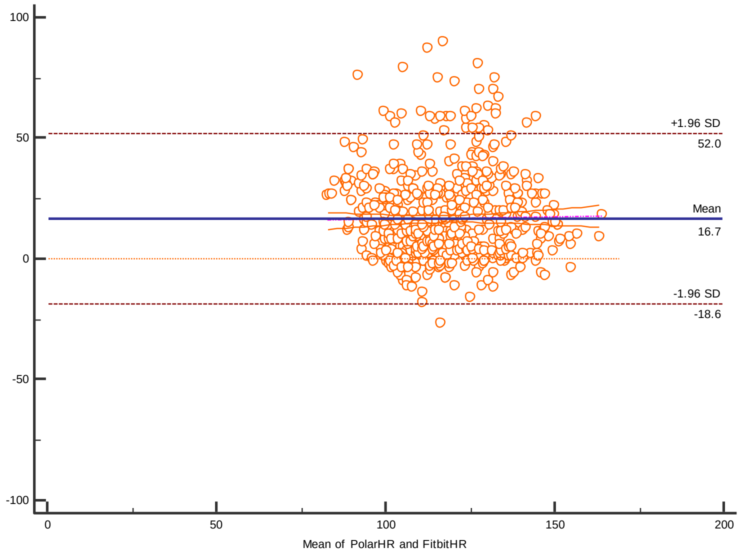
95% CI	1.8046 to 2.0254
Regression Equation	y = -3.5512 + 0.09022 x

C. Light activity HR measurements



95% CI	3.8150 to 4.3085
Regression Equation	y = 13.6664 + -0.1320 x

D. Moderate-Vigorous activity HR measurements



95% CI	15.2617 to 18.1072
Regression Equation	y = 13.9017 + 0.02373 x