Journal of Sports Sciences

ISSN: 0264-0414 (Print) 1466-447X (Online) Journal homepage: http://www.tandfonline.com/loi/rjsp20

Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR

Yang Bai, Paul Hibbing, Constantine Mantis & Gregory J. Welk

To cite this article: Yang Bai, Paul Hibbing, Constantine Mantis & Gregory J. Welk (2017): Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR, Journal of Sports Sciences, DOI: 10.1080/02640414.2017.1412235

To link to this article: https://doi.org/10.1080/02640414.2017.1412235

Published online: 06 Dec 2017.

JOURNAL OF SPORTS SCIENCES, 2017
https://doi.org/10.1080/02640414.2017.1412235

Routledge
Taylor & Francis Group

Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR

Yang Bai [a], Paul Hibbing[b], Constantine Mantis [c] and Gregory J. Welk[c]

[a]Department of Rehabilitation and Movement Science, University of Vermont, Burlington, VT, USA; [b]Department of Kinesiology, Recreation, and Sport Studies, University of Tennessee, Knoxville, TN, USA; [c]Department of Kinesiology, Iowa State University, Ames, IA, USA

ABSTRACT

The purpose of this investigation was to examine the validity of energy expenditure (EE), steps, and heart rate measured with the Apple Watch 1 and Fitbit Charge HR. Thirty-nine healthy adults wore the two monitors while completing a semi-structured activity protocol consisting of 20 minutes of sedentary activity, 25 minutes of aerobic exercise, and 25 minutes of light intensity physical activity. Criterion measures were obtained from an Oxycon Mobile for EE, a pedometer for steps, and a Polar heart rate strap worn on the chest for heart rate. For estimating whole-trial EE, the mean absolute percent error (MAPE) from Fitbit Charge HR (32.9%) was more than twice that of Apple Watch 1 (15.2%). This trend was consistent for the individual conditions. Both monitors accurately assessed steps during aerobic activity ($MAPE_{Apple}$: 6.2%; $MAPE_{Fitbit}$: 9.4%) but overestimated steps in light physical activity. For heart rate, Fitbit Charge HR produced its smallest MAPE in sedentary behaviors (7.2%), followed by aerobic exercise (8.4%), and light activity (10.1%). The Apple Watch 1 had stronger validity than the Fitbit Charge HR for assessing overall EE and steps during aerobic exercise. The Fitbit Charge HR provided heart rate estimates that were statistically equivalent to Polar monitor.

ARTICLE HISTORY
Accepted 24 October 2017

KEYWORDS
Physical activity trackers; energy expenditure; steps; validity; semi-structured protocol

Introduction

The expanding market in wearable activity monitors for consumers has generated considerable interest within the physical activity (PA) research community (https://clinicaltrials.gov/ct2/results?term=activity+tracker&pg=1). Most monitors continue to predict energy expenditure (EE) and step counts, while newer versions of many wrist-worn consumer monitors have also begun incorporating heart rate measures using a technology called photoplethysmography (https://support.apple.com/en-us/HT204666; https://www.fitbit.com/purepulse). This involves flashing green LED light through the skin, to detect the expansion and contraction of wrist capillaries with each pulse (Maeda, Sekine, & Tamura, 2011). Algorithms are then applied to estimate heart rate continuously from that information (Ahmadi, Moradi, Malihi, Karimi, & Shamsollahi, 2015).

The collection of heart rate data from wrist-worn monitors removes the burden of wearing chest straps, and may improve estimates of the intensity and EE of aerobic activities (Achten & Jeukendrup, 2003; Green, 2011), particularly when combined with data from a monitor's internal accelerometer (Brage, Brage, Franks, Ekelund, & Wareham, 2005; Mullan et al., 2015; Spierer, Hagins, Rundle, & Pappas, 2011). However, it is unclear whether heart rate and accelerometer data are being combined to improve EE estimates from consumer monitors, since that information is proprietary. Furthermore, the validity of photoplethysmography itself is not well established (Stahl, An, Dinkel, Noble, & Lee, 2016).

Two market-leading consumer wearable devices are the Apple Watch 1 (Apple Inc., Cupertino, CA USA) and the Fitbit Charge HR (Fitbit Inc., San Francisco, CA USA). Both monitors are worn on the wrist and estimate EE, step count and heart rate via photoplethysmography. However, little is known about the validity of any of these estimates from these two monitors. Thus, the primary purpose of this study is to evaluate the validity of EE from the Apple Watch 1 and the Fitbit Charge HR. A secondary purpose of this study is to assess the accuracy of the raw step count and heart rate estimates produced by these devices.

Methods

Participants

A total of 41 healthy adults aged 19–60 (18 females and 23 males) participated in the study, and data were successfully collected from 39 participants. Recruitment was carried out via email and flyers advertised within the university, as well as through word of mouth. Each participant went through a screening survey. The study was approved by the institutional review board of Iowa State University.

Instruments

Each outcome measure including EE, steps, and heart rate, requires a unique criterion measure. These criterion measures are first described, followed by the descriptions of the monitors and measures being validated.

CONTACT Yang Bai Yang.Bai@med.uvm.edu Department of Rehabilitation and Movement Science, University of Vermont, 310P Rowell Building, Burlington, VT 05405, USA

Criterion measures

The Oxycon Mobile 5.0 (OM, Viasys Healthcare Inc., Yorba Linda, CA) is a small, lightweight (less than 950g) wireless portable metabolic analyzer that was used to perform indirect calorimetry and obtain criterion (minute-by-minute) estimates of EE. Participants wore a facemask and chest-mounted gas analysis and telemetry units. The system recorded and stored breath-by-breath data to capture response to exercise and other physical activities. The OM has been validated against the gold standard Douglas bag method and shown to produce valid results while overcoming limitations of the Douglas bag method (Rosdahl, Gullstrand, Salier-Eriksson, Johansson, & Schantz, 2010). The OM volume and gas was calibrated before each trial.

The Yamax SW-200 DigiWalker (DW) pedometer was used as a criterion measure of steps taken. The Yamax DW has consistently shown accuracy in measuring steps taken at most walking speeds (Bassett et al., 1996; Le Masurier & Tudor-Locke, 2003) and has been widely used for measuring steps in a variety of research applications (Bravata et al., 2007). The pedometer was worn on the waistband anterior to the left iliac crest.

The Polar H7 HR sensor was used as the criterion measure of heart rate. The heart rate strap was placed just below chest muscles and was firmly against the skin. The H7 has been validated in several studies showing high validity compared to ECG (Cheatham, Kolber, & Ernst, 2015). The OM incorporates heart rate telemetry to record the minute-by-minute Polar belt heart rate data as part of its output.

Consumer activity monitors

The present study examines the Apple Watch 1 and the Fitbit Charge HR. Each monitor estimates EE, steps, and heart rate (data not available for Apple Watch 1 in the current study) along with other activity metrics such as distance travelled, active minutes, and sedentary breaks in real time. All outcomes excluding heart rate are reported in cumulative totals. Both of the monitors use photoplethysmography to estimate heart rate from the wrist.

Protocol

Each trial was performed in a large gym setting with sufficient space for each activity. Height and weight were measured using a portable stadiometer and scale, respectively. Waist circumference was measured using a standard tape measure, and percent body fat was assessed using a handheld bioimpedance analysis unit (Omron, Shelton, CT).

Participants were first instructed on how to wear the Polar heart rate strap and the applications for the consumer monitors were initialized to incorporate the participant's demographic and anthropometric information. The Apple Watch 1 and Fitbit Charge HR were positioned on the left wrist and they were then fitted to wear the mask for the OxyCon Mobile.

The 80-minute protocol began with 20 minutes of sedentary activity, performed while seated at a desk. Participants were invited to use a provided laptop, read a book, or use their phone, or similar activities they brought with them. A five-minute break followed this portion, with the next task being 25 minutes aerobic exercise at a self-selected pace walking or jogging on the treadmill. Subjects were allowed to change their pace at any point during this 25-minute session. After another five-minute break, 25-minute simulated free living activities (e.g. folding laundry, sweeping, moving light boxes, stretching, slow walking) followed. Subjects were given minimal direction during this time. The starting and end time of the testing protocol was recorded on the data sheet by researchers.

Data acquisition and processing

The criterion data from the OM were downloaded in minute-by-minute format and included measures of EE and heart rate picked up from the Polar strap. EE values from the OM were summed individually for each of the three components of the protocol, as well as overall for the whole time period. The heart rate data were obtained in minute-by-minute format for both Polar heart rate strap and Fitbit Charge HR, but were analyzed in the same way as EE was (i.e. for the individual conditions as well as the overall trial). Unfortunately, Apple Watch 1 does not automatically measure heart rate at a fixed sampling frequency and the nature of the study design (semi-structured activities) did not allow us to manually obtain heart rate data from Apple Watch 1 without interrupting the activity protocol. Thus, the heart rate validation only applied for the Fitbit Charge HR. The data from the pedometer were obtained by manually recording the accumulated numbers for each segment. Steps were evaluated for the whole 80-minute protocol but also the three segments separately.

The estimates from the monitors were obtained directly from screenshots of the respective applications because the monitors primarily report cumulative totals. The values of EE and steps for individual segments were obtained by subtraction between the respective time periods. For heart rate, minute-by-minute data from the Fitbit Charge HR was accessed through the third-party website Fitabase (Small Steps Labs LLC., San Diego, CA).

Statistical analyses

Participant characteristics (age, weight, height, BMI, waist circumference, and body fat) were summarized using descriptive statistics. Correlational analysis was first carried out using Pearson's correlation coefficient to examine the relationship for each monitor between criterion and estimation. Bland-Altman plots were used to capture the mean difference between criterion and estimation from consumer monitors, with limits of agreement set to mean difference ± 1.96 x standard deviation of mean difference. Mean and standard deviation (SD) of three outcome variables from criterion and consumer monitors were calculated. Individual level measurement errors were evaluated with mean percent errors, mean absolute percent errors (MAPE), and root-mean-square errors. Mean percent error was calculated by averaging the individual percent errors (i.e. (criterion-estimation)/criterion). Mean absolute percent error was calculated by averaging the individual

absolute percent errors (i.e. |(criterion-estimation)/criterion|). Root-mean-square error was calculated as the square root of the mean squared error.

Finally, equivalence testing was performed to directly evaluate the agreement between each measure and the criterion within a 10% range. In equivalence testing, the null and alternative hypotheses are reversed, and the conclusions are subsequently inverted. Consistent with guidelines for equivalence testing, we tested the null hypothesis that there is a difference between the criterion measure and the consumer monitor. In this case, the rejection of the null hypothesis would demonstrate that they are statistically equivalent. Justifying the range of the equivalence region is one of the most difficult aspects of an equivalence test. The equivalence region is typically set based on prior evidence or on the practical meaning of the value. Since there are no definitive guidelines to follow, we adopted the same range of 10% as used in previous studies (Bai et al., 2016; Lee, Kim, & Welk, 2014). Thus, an equivalence zone of ± 10% of the mean of the OM was pre-defined and the statistical test was performed to test if the 90% confidence interval from the consumer monitor measure falls into the defined equivalence zone with 95% precision (α was set up at 0.05). The methodology to test the equivalence null hypothesis has been previously described (Wellek, 2010) and a methodology paper written by our research group is available to provide specific examples (Dixon et al., 2017).

Results

Data from 2 participants was lost because of the OM system failure. Thus, a total of 39 participants (23 males and 16 females) completed the 80-minute sedentary and physical activity protocol. The participant demographic characteristics are summarized in Table 1. A diverse sample was enrolled in the study, with an age range from 19 to 60 and a BMI range of 18.5–37.6. A summary of EE validity indicators from the Apple Watch 1 and the Fitbit Charge HR are reported in Table 2. The Apple Watch 1 slightly underestimated the three sections of the protocol. The Fitbit Charge HR underestimated the sedentary activity and overestimated aerobic exercise and light PA. The mean percent difference for estimates of total activity EE was 7.6% with the Apple Watch 1 but −29.6% with the Fitbit Charge HR. The mean percent difference in all three activities was also correspondingly smaller in magnitude with the Apple Watch 1 compared with the Fitbit Charge HR. The Apple Watch 1 (15.2%) had a whole-trial MAPE that was less than half that of the Fitbit Charge HR (32.9%) and performance was also consistently better for all three separate activity modes in terms of MAPE. Similarly, the RMSE was lower in the Apple Watch 1 than the Fitbit Charge HR for the full routine and the three separate activity modalities. The group level agreement indicator (i.e, Pearson correlation) was higher for the Apple Watch 1 than for the Fitbit Charge HR in the full routine as well as across the three activity modalities: sedentary time ($r_{Apple\ Watch\ 1}$ = 0.45 and $r_{Fitbit\ Charge\ HR}$ = 0.40), aerobic activity ($r_{Apple\ Watch\ 1}$ = 0.88 and $r_{Fitbit\ Charge\ HR}$ = 0.73), and light PA ($r_{Apple\ Watch\ 1}$ = 0.66 and $r_{Fitbit\ Charge\ HR}$ = 0.41). The Bland-Altman plots based on EE from 80-minute protocol revealed less systematic bias from the Apple Watch 1 than with the Fitbit Charge HR, except for three participants who had extremely high errors (Figure 1). While the estimates were better for the Apple Watch 1, only the estimated EE in light PA sessions with this device fell in the equivalence testing zone (Table 2).

The validity of step estimation from the Apple Watch 1 and the Fitbit Charge HR is summarized in Table 3. Both of the monitors overestimated the steps compared to the Yamax DW for 80-minute routine and the overestimation was primarily attributable to light PA. The Apple Watch 1 had narrower 95% confidence internal of mean bias than the Fitbit Charge HR but both the Apple Watch 1 and the Fitbit Charge HR had relatively well-distributed data. Both consumer monitors had low measurement errors in estimating steps during the aerobic workout phase. The mean percent differences were 0.7% from the Apple Watch 1 and 4.5% from the Fitbit Charge HR, and the MAPE was less than 10% for both the Apple Watch 1 (6.2%) and the Fitbit Charge HR (9.4%). The correlations between the consumer monitors and the Yamax DW were also higher in the aerobic exercise phase ($r_{Apple\ Watch\ 1}$ = 0.91 and $r_{Fitbit\ Charge\ HR}$ = 0.86) than during light PA ($r_{Apple\ Watch\ 1}$ = 0.30 and $r_{Fitbit\ Charge\ HR}$ = 0.32) or the overall 80-minute routine ($r_{Apple\ Watch\ 1}$ = 0.79 and $r_{Fitbit\ Charge\ HR}$ = 0.77). Similarly, the measurement errors were much higher during the light PA session. Both the Apple Watch 1 and the Fitbit Charge HR overestimated steps with respective average estimates of 827 and 1,002 steps during light PA compared to a mean of 571 steps from the Yamax DW. Steps assessed by the Apple Watch 1 were statistically equivalent to the Yamax DW (within 10% range) for the aerobic portion and the total 80-minute protocol (Figure 2). The step counts captured by the Fitbit Charge HR in the aerobic workout were marginally in the equivalence zone (Table 3).

The overall agreement of heart rate measurement between Fitbit Charge HR and the criterion measure Polar belt was high across all three activity segments (Table 4). The Fitbit Charge HR underestimated the heart rate during sedentary activity, which resulted in a mean difference of 2.3%, MAPE of 7.2%, and RMSE of 7.2 beats per minute. The Fitbit Charge HR also underestimated the aerobic exercise and light PA with MAPE values of 8.4% and 10.1%, respectively. The Fitbit Charge HR had a higher correlation with measured heart rate during the sedentary activity segment (r = 0.85) and the aerobic segment (r = 0.84) compared with light PA (r = 0.74). In the Bland-Altman plots, the Fitbit Charge HR had less evidence of systematic bias in estimating aerobic activity heart rate compared to the other conditions (Figure 1). The Fitbit Charge HR tended to

Table 1. Descriptive demographics of the participants.

	Mean	SD	Minimum	Maximum
Age (yr)	32.0	11.0	19.0	60.0
Height (cm)	171.7	9.5	156.0	197.0
Weight (kg)	72.7	12.8	48.0	100.6
BMI(kg/m^2)	24.7	4.0	18.5	37.6
BF (%)	24.3	13.2	4.7	41.3
Waist (cm)	83.2	9.8	64.5	109.0

Abbreviations: BMI, Body Mass Index; BF, Body Fat Percentage; SD, Standard Deviation.

Table 2. Validity of energy expenditure estimation from Apple Watch and Fitbit Charge HR.

	Mean (SD)	MPE (SD)	MAPE (SD)	RMSE	Correlation	Equivalence Zone
Oxycon Mobile						
Total	317.5 (103.3)					(286.2, 349.8)
SB	40.0 (10.2)					(36.0, 44.0)
AE	191.8 (83.3)					(172.6, 211.0)
LPA	86.4 (23.2)					(77.5, 94.8)
Apple Watch						
Total	287.4 (112.5)	7.6% (19.7%)	15.2% (14.5%)	56.2	0.89**	(261.0, 315.2)
SB	36.0 (10.4)	9.4% (26.1%)	22.2% (17.8%)	11.4	0.45**	(32.6, 37.2)
AE	168.7 (79.9)	6.3% (24.3%)	18.2% (17.4%)	44.7	0.88**	(151.2, 190.9)
LPA	82.5 (34.5)	3.7% (32.2%)	23.4% (22.1%)	24.2	0.66**	(73.2, 90.9)
Fitbit Charge HR						
Total	387.6 (136.1)	−29.6% (33.4%)	32.9% (30.0%)	117.9	0.75**	(363.4, 434.8)
SB	31.0 (18.5)	21.3% (28.1%)	34.6% (27.4%)	19.7	0.4*	(26.2, 36.4)
AE	223.2 (84.0)	−32.3% (45.1%)	41.4% (37.6%)	71.3	0.73**	(209.1, 252.9)
LPA	133.3 (59.7)	−60.3% (63.1%)	61.2% (62.2%)	74.5	0.41*	(120.1, 152.9)

Abbreviations: AE, Aerobic Exercise; LPA, Light Physical Activity; MAPE, Mean Absolute Percent Error; MPE, Mean Percent Error; RMSE, Root Mean Square Error; SB, Sedentary Behavior; SD, Standard Deviation.
Energy expenditure was expressed in kcals.
** Correlation is significant at an alpha level of 0.01 (2-tailed). *Correlation is significant at an alpha level of 0.05 (2-tailed).

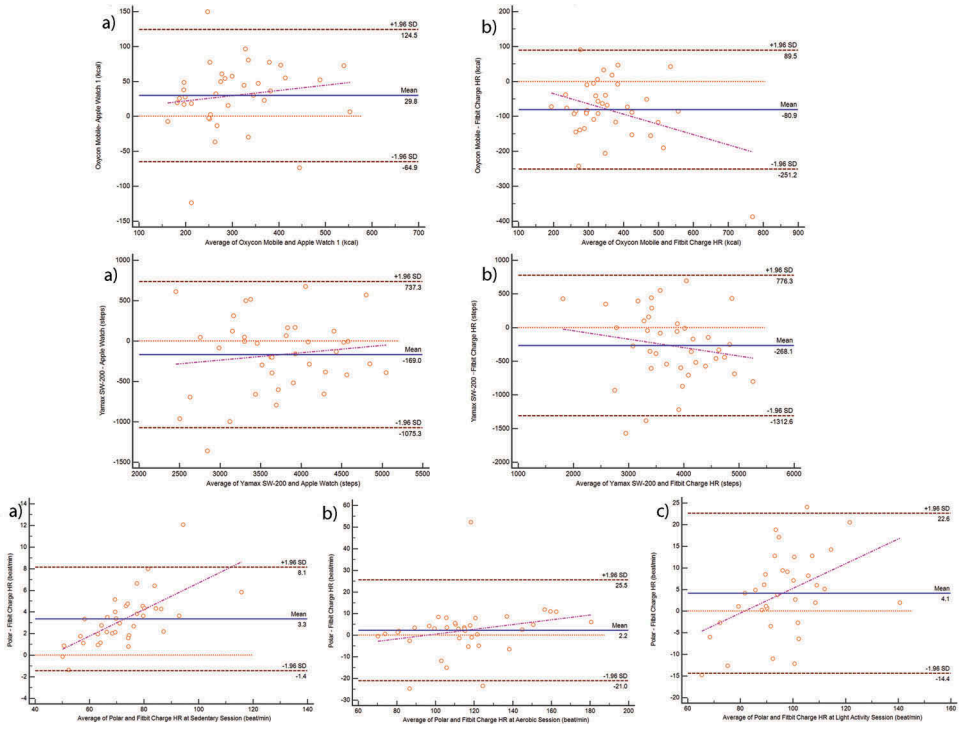

Figure 1. Bland-Altman plots comparing energy expenditure, steps, and heart rate between criterion measures and two consumer activity monitor Apple Watch 1 and Fitbit Charge HR.

Note: top panel indicates the results for total energy expenditure from two consumer monitors; middle panel indicates the results for total step counts from two consumer monitors; bottom panel indicates the results for heart rate from Fitbit Charge HR separately for the three activity sessions.

underestimate heart rate in higher intensity, light activity and participants with higher sedentary heart rate. The heart rate estimated during sedentary, aerobic and light PA sessions all fell in the designated equivalence zone (Table 4 and Figure 2).

Discussion

The results of this study showed that the Apple Watch 1 estimated EE with relative accuracy, but the Fitbit Charge HR did not. Both monitors had a high validity for assessing steps during aerobic activity, and the Fitbit Charge HR accurately predicted heart rate. The present study is among the first to systematically evaluate the accuracy of consumer physical activity monitors with built-in heart rate. It provides evidence to both scholars and consumers on how accurate these newer device models are.

The current study found reasonable validity for Apple Watch 1 since the overall MAPE for estimating EE (15.2%) was comparable with estimates from the SenseWear armband (15.3%) included in a previous study with a similar activity protocol (Bai et al., 2016). However, the results somehow contrast with a study by Wallen et al. (Wallen et al., 2016) that recently evaluated the Apple Watch 1. They reported extremely low correlations with indirect calorimetry (r = 0.16) while we observed strong correlations

Table 3. Validity of steps estimation from Apple Watch and Fitbit Charge HR.

	Mean (SD)	MPE (SD)	MAPE (SD)	RMSE	Correlation	Equivalence Zone
Yamax SW-200	Pedometer					
Total	3637 (723)					(3277.6, 4006.0)
SB	13 (37)					(11.9, 14.5)
AE	3044 (616)					(2740.0, 3348.9)
LPA	571 (266)					(514.3, 628.6)
Apple Watch						
Total	3815 (646)	−6.4% (16.9%)	11.8% (13.5%)	487	0.79**	(3621.2, 3985.8)*
SB	13 (39)	−271.6% (1691.6%)[a]	452.9% (1743.3%)[a]	20[a]	0.87**	(3.4, 23.8)
AE	2975 (610)	0.7% (11.7%)	6.2% (10.1%)	272	0.91**	(2841.5, 3171.6)*
LPA	827 (322)	148.4% (515.4%)[a]	160.8% (511.6%)[a]	382[a]	0.3	(698.7, 849.8)
Fitbit Charge HR						
Total	3911 (788)	−8.5% (18.7%)	14.3% (14.6%)	590	0.77**	(3689.3, 4130.6)
SB	16 (36)	530.0% (233.1%)[a]	595.9% (2391.7%)[a]	28[a]	0.71**	(6.5, 25.6)
AE	2891 (702)	4.5% (15.9%)	9.4% (13.4%)	372	0.86**	(2719.7, 3098.1)
LPA	1002 (357)	222.5% (750.0%)[a]	231.0% (7.59%)[a]	532[a]	0.32*	(873.7, 1057.0)

Abbreviations: AE, Aerobic Exercise; LPA, Light Physical Activity; MAPE, Mean Absolute Percent Error; MPE, Mean Percent Error; RMSE, Root Mean Square Error; SB, Sedentary Behavior; SD, Standard Deviation.

** Correlation is significant at an alpha level of 0.01 (2-tailed). *Correlation is significant at an alpha level of 0.05 (2-tailed).

[a] The errors were largely inflated because the very limited steps accumulated during the sedentary and light physical activity. For instance, if participants walked 5 steps but the monitor recorded 15, the MAPE would be 200%.

Table 4. Validity of heart rate estimation from Fitbit Charge HR.

	Mean (SD)	MPE (SD)	MAPE (SD)	RMSE	Correlation	Equivalence Zone
Polar HR Sensor						
SB	72.8 (12.4)					(65.9, 80.5)
AE	116.6 (31.0)					(104.3, 127.4)
LPA	97.1 (18.6)					(87.7, 107.2)
Fitbit Charge HR						
SB	70.0 (12.0)	2.3% (8.9%)	7.2% (7.5%)	7.2	0.85**	(67.6, 74.9)*
AE	115.9 (30.6)	−0.2% (8.3%)	8.4% (7.1%)	17.8	0.84**	(108.2, 122.1)*
LPA	93.9 (15.7)	2.2% (10.0%)	10.1% (5.1%)	13.2	0.74**	(89.9, 97.9)*

Abbreviations: AE, Aerobic Exercise; HR, Heart Rate; LPA, Light Physical Activity; MAPE, Mean Absolute Percent Error; MPE, Mean Percent Error; RMSE, Root Mean Square Error; SB, Sedentary Behavior; SD, Standard Deviation.

** Correlation is significant at an alpha level of 0.01 (2-tailed). *Correlation is significant at an alpha level of 0.05 (2-tailed).

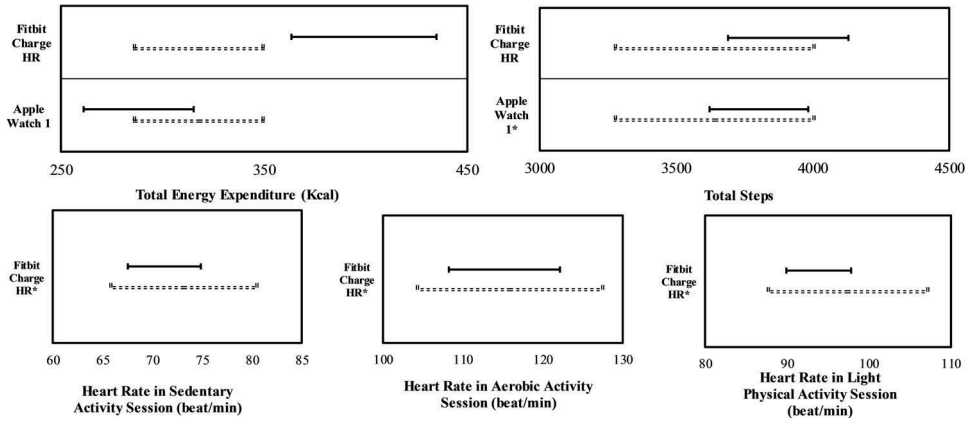

Figure 2. Agreement from 95% equivalence testing between criterion measures and two consumer activity monitor Apple Watch 1 and Fitbit Charge HR.

Note: Dash lines indicate the equivalence zone from criterion measure; Dark lines indicate the 90% confidence interval of the estimated from activity monitors.* Within the equivalence zone.

(r = 0.86) and lower measurement errors. Wallen et al. reported a large underestimation of EE from the Apple Watch 1 (mean of 162.6 kcal) compared with indirect calorimetry (mean of 285.7 kcal), but we observed a relatively low mean error of 30.1 kcal (and low overall MAPE) between the Apple Watch 1 and indirect calorimetry. The validity of Apple Watch from the current study is

comparable to previous consumer monitors validation studies reporting high correlation, and MAPE between 15% to 10%. Considering the high validity of heart rate measures in Apple Watch from the Shcherbina et al. study (Shcherbina et al., 2016) and Wallen's study, it is surprising that Wallen's study showed such low validation in EE estimation from Apple Watch. No evident clue could explain the substantial difference between the two studies, but the divergent results certainly require further investigation.

The methods used internally in new consumer monitors are not fully disclosed so it is not clear how the estimates of EE are derived. The description from the Apple website states that the Apple Watch 1 self-selects the most appropriate inputs (e.g., accelerometer, heart rate, GPS) to measure activity depending on the type of activity detected (e.g., indoor running, outdoor cycling, walking). This implies that the heart rate is directly used in establishing or refining the estimate but it is not clear if this data channel is used in a similar way (or not) within the Fitbit. Consistent with past studies (Bai et al., 2016), performance of monitors tends to be better for evaluating overall 80-minute routine compared to EE accuracy in three separate sessions since it enables some degree of cancelation of over- and underestimation from different activity modalities. The Apple Watch 1 performed the best during aerobic exercise, but it underestimated EE in all three activity segments.

The Fitbit Charge HR had much lower validity for estimating EE compared with the Apple Watch 1. A number of past studies have provided support for the relatively good performance of earlier Fitbit products, such as Fitbit One™ (Lauritzen, Munoz, Luis Sevillano, & Civit, 2013), Fitbit Zip™(Lee et al., 2014), and Fitbit Flex™ (Bai et al., 2016), none of which has the built-in heart rate feature. The MAPE was 10.4%, 10.1%, and 16.8%, respectively, for the earlier three Fitbit models in estimating EE against indirect calorimetry, in spite of the study design and placement of monitors (One™ and Zip™ are worn on the waist) were different between the current and previous studies. The MAPE from the present study was more than twice that from the previous validation study so it is hard to explain this relatively poor performance. Shcherbina et al. tested Fitbit Surge and Apple Watch and found lower error in Fitbit Surge compared to Apple Watch for aerobic activity dominated protocol. The overall errors remained high in Fitbit Surge (i.e., median 27.4% mean percent error) for estimating EE (Shcherbina et al., 2016). One of the possible explanations could be that the EE prediction algorithms in this new device were not fully developed, as we tested the monitor right after it was released on the market. Our study along with other two recent studies all found high validity in heart rate estimation but relatively low validity in measuring EE for Fitbit products (i.e., Surge and Charge 2) (Shcherbina et al., 2016; Wallen et al., 2016). It is unknown whether the algorithms in the Fitbit Charge HR were even updated or whether efforts were made to directly integrate and use the heart rate data in the estimations. This proprietary information is unavailable, further demonstrating the inherent "Black Box" conundrum that continues to broadly affect efforts to study performance of accelerometry-based consumer monitors.

While the EE estimates were poor, it is noteworthy that the Fitbit Charge HR showed relatively high validity in heart rate estimation, with estimates ranging from 7.2% to 10.1% in MAPE for sedentary, aerobic exercise, and light PA. These results are comparable to a recent validation study by Stahl et al., which evaluated the heart rate estimation from Fitbit Charge HR and found a range of 1.73% to 10.06% MAPE across rest, cool down and locomotive activity at speeds ranging from 3.2 km/h to 9.6 km/h (Stahl et al., 2016). Other studies have also examined the validity of other less popular commercially available activity monitors with photopletysmographic heart rate feature such as Mio Alpha, Omron HR500U, and Schosche Rhythm. A relatively small MAPE has been reported in another study (Spierer, Rosen, Litman, & Fujii, 2015), but the overall range was close to what we observed. It is unclear whether systematic bias occurs in measuring heart rate at different intensities (i.e. higher vs. lower heart rate) or activity types. Although Fitbit has faced a class-action lawsuit in which the plaintiffs claimed the heart rate measure was inaccurate, the lawsuit was based on an unpublished study that used questionable statistical approaches.

Minute-by-minute heart rate data were unavailable from the Apple Watch 1 (or the Health application on iphone) and could not be directly assessed in this study. However, Wallen et al. (Wallen et al., 2016) reported strong accuracy for measuring heart rate with correlations of 0.95 and a mean difference of 1.3 beats/minute compared to ECG. The authors reported manually measuring the heart rate from Apple Watch 1 on the wrist every 15 seconds but they only reported one average heart rate from the full 58-minute protocol. The lack of continuous recording of HR complicates evaluation of these features in consumer monitors so additional research is warranted.

For steps, both monitors performed more similarly. The validity of both the Apple Watch 1 and the Fitbit Charge HR were reasonable for aerobic exercise but much less accurate for light PA with patterns of consistent overestimation in this phase. Overestimation for light activity is somewhat expected considering the monitor is designed to be worn on the wrist, and the activities designed in the light PA sessions (e.g. folding laundry, sweeping, moving light boxes, stretching, slow walking) were largely non-locomotion activities that still involved arm movements. The large overall measurement errors for light PA are not surprising because it has been well documented that error in pedometers is considerably larger for low intensity locomotive activities (Crouter, Schneider, Karabulut, & Bassett, 2003; Lee, Williams, Brown, & Laurson, 2015; Schneider, Crouter, & Bassett, 2004). Recent studies have also documented weaker performance for low intensity household activities such as vacuuming, dusting, filing papers, cleaning rooms, folding laundry, pushing a stroller, etc. (Chen, Kuo, Pellegrini, & Hsu, 2016; Hickey, John, Sasaki, Mavilia, & Freedson, 2016). A recent study conducted by Chen et al. found similar results as our study did for Fitbit flex (Chen et al., 2016). Several other studies have reported that wrist-worn consumer activity monitors such as the Jawbone (Ferguson, Rowlands, Olds, & Maher, 2015) and research grade monitors such as the ActiGraph also

overestimate steps when worn on the wrist (Tudor-Locke, Barreira, & Schuna, 2015).

While the study provides novel insights about these new heart rate monitors, some limitations should be noted. Specifically, heart rate data could not be assessed from the Apple Watch 1. The lack of minute-by-minute downloadable heart rate data from the Apple Watch 1 was essentially unavoidable, because of the lack of flexibility offered by the application. Additionally, steps data was mainly discussed at the whole-trial level and aerobic activity. The focus on total steps and steps during aerobic activity was more appropriate for the present study because we intentionally included sedentary and light activities in the design to validate the EE and heart rate, but steps are not relevant for these types of conditions. Tracking of steps for low intensity activities is a known limitation for all pedometers (including the Yamax DW) so it was deemed more important to evaluate patterns over the full 80 minutes. The placement of the monitors (both on left wrist) was fixed because some activities such as writing and sweeping involve one arm more than the other. However, users could choose to wear it on their dominant or non-dominant arms. It is possible that placement of the monitors could potentially influence the validity of heart rate measurement. Counterbalanced placement could be used in future studies to minimize any potential bias caused by the placement in the current study. Another source of random error could be from the food thermogenesis because the participants were not asked to come in a fasting state. Most of the participants did not perform vigorous physical activities before participating rather than active commute (e.g., walking and biking) to the lab from some participants.

The study provided a real-world evaluation of the monitors by including free-living behaviors across a range of intensities including sedentary and light PA. The relatively high measurement errors may be attributable to the latitude given to participants to select comfortable activities during the testing period, particularly aerobic and light PA portion of the test. However, this format theoretically mimics everyday activity more effectively than traditional structured activities. Thus, it may reflect more realistic estimates of validity than other protocols. The formal evaluation of the photoplethysmographic heart rate estimates from the Fitbit Charge HR also provides new evidence of how effective this novel technique is when used in consumer devices. Future studies should evaluate the accuracy of moderate and vigorous PA time estimates from consumer monitors, since this is a vital indicator in health-related research. Similarly, the accuracy of activity recognition features from several new monitors including the Apple Watch 1 and the Fitbit Charge HR, which automatically classify bouts activity into walk, run, biking, aerobic workout etc., needs further assessment.

Acknowledgments

We acknowledge the specific contributions of undergraduate research assistants Sydney Reeves and Matthew Harm who contributed significantly during data collection. None of the authors have a professional relationship with companies or manufacturers who might benefit from the results of the present study. The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. There is no funding supported the data collection and writing the manuscript. The authors declare that they have no competing interests.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Yang Bai http://orcid.org/0000-0001-6751-3896
Constantine Mantis http://orcid.org/0000-0003-3897-3636

References

Achten, J., & Jeukendrup, A. E. (2003). Heart rate monitoring: Applications and limitations. *Sports Medicine*, *33*(7), 517–538. doi:10.2165/00007256-200333070-00004

Ahmadi, A. K., Moradi, P., Malihi, M., Karimi, S., & Shamsollahi, M. B. (2015). Heart Rate monitoring during physical exercise using wrist-type photoplethysmographic (PPG) signals. *Conference Proceedings IEEE Engineering in Medicine and Biology Society*, *(2015*, 6166–6169. doi:10.1109/embc.2015.7319800

Bai, Y., Welk, G. J., Nam, Y. H., Lee, J. A., Lee, J.-M., Kim, Y., … Dixon, P. M. (2016). Comparison of consumer and research monitors under semi-structured settings. *Medicine & Science in Sports & Exercise*, *48*(1), 151–158. doi:10.1249/mss.0000000000000727

Bassett, D. R., Jr., Ainsworth, B. E., Leggett, S. R., Mathien, C. A., Main, J. A., Hunter, D. C., & Duncan, G. E. (1996). Accuracy of five electronic pedometers for measuring distance walked. *Medicine & Science in Sports & Exercise*, *28*(8), 1071–1077. doi:10.1097/00005768-199608000-00019

Brage, S., Brage, N., Franks, P. W., Ekelund, U., & Wareham, N. J. (2005). Reliability and validity of the combined heart rate and movement sensor Actiheart. *European Journal of Clinical Nutrition*, *59*(4), 561–570. doi:10.1038/sj.ejcn.1602118

Bravata, D. M., Smith-Spangler, C., Sundaram, V., Gienger, A. L., Lin, N., Lewis, R., … Sirard, J. R. (2007). Using pedometers to increase physical activity and improve health: A systematic review. *JAMA*, *298*(19), 2296–2304. doi:10.1001/jama.298.19.2296

Cheatham, S. W., Kolber, M. J., & Ernst, M. P. (2015). Concurrent validity of resting pulse-rate measurements: A comparison of 2 smartphone applications, the polar H7 belt monitor, and a pulse oximeter with bluetooth. *Journal of Sport Rehabilitation*, *24*(2), 171–178. doi:10.1123/jsr.2013-0145

Chen, M.-D., Kuo, -C.-C., Pellegrini, C. A., & Hsu, M.-J. (2016). Accuracy of wristband activity monitors during ambulation and activities. *Medicine & Science in Sports & Exercise*, *48*(10), 1942–1949. doi:10.1249/mss.0000000000000984

Crouter, S. E., Schneider, P. L., Karabulut, M., & Bassett, D. R., Jr. (2003). Validity of 10 electronic pedometers for measuring steps, distance, and energy cost. *Medicine & Science in Sports & Exercise*, *35*(8), 1455–1460. doi:10.1249/01.mss.0000078932.61440.a2

Dixon, P. M., Saint-Maurice, P. F., Kim, Y., Hibbing, P., Bai, Y., & Welk, G. J. (2017). A primer on the use of equivalence testing for evaluating measurement agreement. *Medicine and Science in Sports and Exercise*. Advance online publication. doi:10.1249/MSS.0000000000001481

Ferguson, T., Rowlands, A. V., Olds, T., & Maher, C. (2015). The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: A cross-sectional study. *The International Journal of Behavioral Nutrition and Physical Activity*, *12*, 42. doi:10.1186/s12966-015-0201-9

Green, J. A. (2011). The heart rate method for estimating metabolic rate: Review and recommendations. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *158*(3), 287–304. doi:10.1016/j.cbpa.2010.09.011

Hickey, A., John, D., Sasaki, J. E., Mavilia, M., & Freedson, P. (2016). Validity of activity monitor step detection is related to movement patterns. *Journal of Physical Activity and Health*, *13*(2), 145–153. doi:10.1123/jpah.2015-0203

Lauritzen, J., Munoz, A., Luis Sevillano, J., & Civit, A. (2013). The usefulness of activity trackers in elderly with reduced mobility: A case study. *Studies in Health Technology and Informatics*, *192*, 759–762.

Le Masurier, G. C., & Tudor-Locke, C. (2003). Comparison of pedometer and accelerometer accuracy under controlled conditions. *Medicine & Science in Sports & Exercise*, *35*(5), 867–871. doi:10.1249/01.mss.0000064996.63632.10

Lee, J. A., Williams, S. M., Brown, D. D., & Laurson, K. R. (2015). Concurrent validation of the Actigraph gt3x+, Polar Active accelerometer, Omron HJ-720 and Yamax Digiwalker SW-701 pedometer step counts in lab-based and free-living settings. *Journal of Sports Sciences*, *33*(10), 991–1000. doi:10.1080/02640414.2014.981848

Lee, J.-M., Kim, Y., & Welk, G. J. (2014). Validity of consumer-based physical activity monitors. *Medicine & Science in Sports & Exercise*, *46*(9), 1840–1848. doi:10.1249/mss.0000000000000287

Maeda, Y., Sekine, M., & Tamura, T. (2011). The advantages of wearable green reflected photoplethysmography. *Journal of Medical Systems*, *35*(5), 829–834. doi:10.1007/s10916-010-9506-z

Mullan, P., Kanzler, C. M., Lorch, B., Schroeder, L., Winkler, L., Laich, L., … Pasluosta, C. (2015). Unobtrusive heart rate estimation during physical exercise using photoplethysmographic and acceleration data. *Conference Proceedings IEEE Engineering in Medicine and Biology Society*, *2015*, 6114–6117. doi:10.1109/embc.2015.7319787

Rosdahl, H., Gullstrand, L., Salier-Eriksson, J., Johansson, P., & Schantz, P. (2010). Evaluation of the Oxycon Mobile metabolic system against the Douglas bag method. *European Journal of Applied Physiology*, *109*(2), 159–171. doi:10.1007/s00421-009-1326-9

Schneider, P. L., Crouter, S., & Bassett, D. R. (2004). Pedometer measures of free-living physical activity: Comparison of 13 models. *Medicine & Science in Sports & Exercise*, *36*(2), 331–335. doi:10.1249/01.mss.0000113486.60548.e9

Shcherbina, A., Mattsson, C. M., Waggott, D., Salisbury, H., Christle, J. W., Hastie, T. J., … Ashley, E. A. (2016). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *bioRxiv*. doi:10.1101/094862

Spierer, D. K., Hagins, M., Rundle, A., & Pappas, E. (2011). A comparison of energy expenditure estimates from the Actiheart and Actical physical activity monitors during low intensity activities, walking, and jogging. *European Journal of Applied Physiology*, *111*(4), 659–667. doi:10.1007/s00421-010-1672-7

Spierer, D. K., Rosen, Z., Litman, L. L., & Fujii, K. (2015). Validation of photoplethysmography as a method to detect heart rate during rest and exercise. *Journal of Medical Engineering & Technology*, *39*(5), 264–271. doi:10.3109/03091902.2015.1047536

Stahl, S. E., An, H.-S., Dinkel, D. M., Noble, J. M., & Lee, J.-M. (2016). How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport & Exercise Medicine*, *2*(1), e000106. doi:10.1136/bmjsem-2015-000106

Tudor-Locke, C., Barreira, T. V., & Schuna, J. M., Jr. (2015). Comparison of step outputs for waist and wrist accelerometer attachment sites. *Medicine & Science in Sports & Exercise*, *47*(4), 839–842. doi:10.1249/mss.0000000000000476

Wallen, M. P., Gomersall, S. R., Keating, S. E., Wisløff, U., Coombes, J. S., & Calbet, J. A. L. (2016). Accuracy of heart rate watches: Implications for weight management. *PLoS One*, *11*(5), e0154420. doi:10.1371/journal.pone.0154420

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed). Boca Raton, FL: CRC Press.