ORIGINAL PAPER

The validity of two widely used commercial and research-grade activity monitors, during resting, household and activity behaviours

R. O'Driscoll[1] · J. Turicchi[1] · M. Hopkins[2] · C. Gibbons[1] · S. C. Larsen[3] · A. L. Palmeira[4,5] · B. L. Heitmann[3,6,7] · G. W. Horgan[8] · G. Finlayson[1] · R. J. Stubbs[1]

Abstract

Wearable devices are increasingly prevalent in research environments for the estimation of energy expenditure (EE) and heart rate (HR). The aim of this study was to validate the HR and EE estimates of the Fitbit charge 2 (FC2), and the EE estimates of the Sensewear armband mini (SWA). We recruited 59 healthy adults to participate in walking, running, cycling, sedentary and household tasks. Estimates of HR from the FC2 were compared to a HR chest strap (Polar) and EE to a stationary metabolic cart (Vyntus CPX). The SWA overestimated overall EE by 0.03 kcal/min^{-1} and was statistically equivalent to the criterion measure, with a mean absolute percentage error (MAPE) of 29%. In contrast, the FC2 was not equivalent overall (MAPE = 44%). In household tasks, MAPE values of 93% and 83% were observed for the FC2 and SWA, respectively. The FC2 HR estimates were equivalent to the criterion measure overall. The SWA is more accurate than the commercial-grade FC2. Neither device is consistently accurate across the range of activities used in this study. The HR data obtained from the FC2 is more accurate than its EE estimates and future research may focus more on this variable.

Keywords Energy Expenditure · Heart Rate · Wearables · Accelerometer · Validation

Abbreviations

ANOVA	Analysis of variance
BPM	Beats per minute
DBP	Diastolic blood pressure
EE	Energy expenditure
FC2	Fitbit Charge 2
FM	Fat mass
FFM	Fat-free mass
HR	Heart rate
MAE	Mean absolute error
MAPE	Mean absolute percentage error
PA	Physical activity
RMR	Resting metabolic rate
RHR	Resting heart rate
RMSE	Root mean squared error
SWA	SenseWear Armband Mini
SBP	Systolic blood pressure

✉ R. O'Driscoll
Psrod@leeds.ac.uk

1. Appetite Control and Energy Balance Group, School of Psychology, University of Leeds, Leeds, UK

2. School of Food Science and Nutrition, Faculty of Mathematics and Physical Sciences, University of Leeds, Leeds, UK

3. Research Unit for Dietary Studies, The Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, The Capital Region, Denmark

4. Faculdade de Motricidade Humana, Universidade de Lisboa, Lisbon, Portugal

5. Universidade Lusófona, Lisbon, Portugal

6. Section for General Medicine, Department of Public Health, Copenhagen University, Copenhagen, Denmark

7. The Boden Institute, Charles Perkins Centre, University of Sydney, Sydney, Australia

8. Biomathematics & Statistics Scotland, Aberdeen, UK

1 Introduction

An increased participation in physical activity (PA) and a more active lifestyle is associated with a reduced risk of obesity and prevention of weight regain following weight loss [1–6]. Increases in PA can not only elevate energy expenditure (EE), but also influence the control of appetite and energy

intake [7]. Thus, the quantification of PA and EE represent primary areas of interest in the study appetite and energy balance. Wearable devices, relying primarily on accelerometery, have been available for the assessment of PA and EE in research environments for some time [8–10]. Commercial-grade wearable devices are increasingly used in large-scale PA and dietary research, but their use in such environments is dependent on their ability to accurately and precisely track and estimate the energy cost of a wide range of activities.

The ability to estimate EE using cost effective and practical wearable devices has long been of scientific interest [11–13] as such devices would help overcome limitations associated with currently available techniques. For example, indirect calorimetry methods are generally limited to laboratory environments and expensive stable isotopic criterion techniques provide mean estimates of daily EE over 10–14 days and do not capture daily variation in EE [14]. These issues constrain their use in large-scale research and limit their utility for the collection of continuous EE data over long-term periods of time in free-living individuals. Accurate estimates of EE from discrete wearable devices would add a new dimension to the assessment of free-living EE across a range of activities and population groups in health and disease. Recent developments in wearable technology and cloud storage capacity means it is now theoretically possible and practical to continuously monitor EE patterns in the free-living individual [15]. However, inaccurate instruments are undesirable as they may bias interpretation of data outcomes [16].

A body of literature validating wearable devices exists [17, 18] but product release is often faster than validation studies [19] and thus, the accuracy of newer devices remains uncertain. Physiological sensors, including heart rate (HR) sensors [20] are commonplace in newer activity monitors [21] and such innovation may be bringing the accuracy of commercial devices in line with more established research-grade devices [22]. A linear relationship exists between oxygen consumption (VO_2) and HR during moderate to high intensity activities [23, 24] and therefore monitoring HR at the minute-level enables relative PA intensity [25, 26] or EE [27] to be estimated. It seems that combination approaches, in which physiological and movement variables are incorporated into predictive algorithms, improves the estimation of PA or EE relative to accelerometery alone [21, 28]. For HR to be used to monitor PA or EE in wearable activity monitors it is imperative that HR estimates are valid in populations and activities of interest.

There is considerable interest in measuring HR and EE with accuracy and precision in research, clinical and consumer environments. The purpose of the present study is to evaluate the validity the HR and EE estimates of the Fitbit Charge 2 (FC2), a modern commercial grade wearable device and the EE estimates of the research-grade SenseWear Armband Mini (SWA) during sedentary, household, ambulatory and cycling tasks in a heterogeneous population.

2 Methods

2.1 Participants

A diverse sample ($n = 59$) was enrolled in the study (age range: 22–73 years, weight range 49.2–105.99 kg) and participant characteristics are presented in Table 1. Participants were primarily recruited from the Leeds centre of the NoHoW trial ($n = 44$), a randomized controlled trial testing the efficacy of an ICT based toolkit for weight loss maintenance across three European centres: United Kingdom, (Leeds), Denmark (Copenhagen), and Portugal (Lisbon). The main trial is registered with the ISRCTN registry (ISRCTN88405328). Participants recruited from the NoHoW trial were provided with their own FC2. In addition, 15 participants were recruited from the local area. Exclusion criteria for the present study included: pregnancy, medications associated with alteration to metabolic rate, the inability to ambulate without assistance, the presence or sign of cardiovascular, metabolic, renal disorders, illness or injury that provide an increased risk of medical events during PA [29]. This study was conducted at the Appetite Control and Energy Balance research laboratory at The University of Leeds, and participants provided written informed consent for this specific study prior to participation. The experimental protocol was approved by The University of Leeds, School of Psychology ethics committee (PSC-407, 18/08/2018).

2.2 Study protocol

Following body composition and RMR measurements (described below), participants transitioned to the exercise laboratory where the PA protocol was performed. Participants were initially seated for 5 min, followed by 5 min standing. Next participants performed 5 min of treadmill walking (4 km/h), incline walking (4 km/h, 5% incline), running (6–8 km/h, 5% incline) and incline running (6–8 km/h, 5% incline). Participants were then given a 3-min resting period and then transitioned to a cycle ergometer and performed 5 min of low-intensity (30 watts), and moderate intensity cycling (60 watts). Lastly, after another resting period, participants performed a 5-min folding task and a 5-min sweeping task. Throughout this protocol, participants wore a polar HR monitor, FC2 and a SWA at all times whilst breath by breath respiratory data was collected using a stationary metabolic cart.

2.3 Physical measurements

Participants arrived at the laboratory in a fasted state having refrained from the intake of food, caffeine and exercise in the 12 h prior to testing. After completing a medical screening questionnaire and providing informed consent, height was

Table 1 Characteristics of the participants

	N	Age	Height	Weight	FFM%	FM%	FM (kg)	FFM (kg)	RMR (kcal)	SBP	DBP	RHR
	59	44.42 ± 14.14	167.45 ± 8.91	75.65 ± 13.63	67.54 ± 10.31	32.46 ± 10.31	24.83 ± 10.73	49.81 ± 8.85	1581.82 ± 280.43	121.92 ± 11.46	78.05 ± 8.54	64.93 ± 10
F	41	46.56 ± 13.09	163.10 ± 6.23	71.47 ± 12.92	64.42 ± 8.76	35.58 ± 8.76	26.29 ± 10.59	45.17 ± 4.78	1466.60 ± 223.90	118.44 ± 11.51	77.37 ± 8.99	67.07 ± 9.96
M	18	39.79 ± 15.54	176.85 ± 6.11	84.69 ± 10.57	75.52 ± 9.89	24.48 ± 9.89	21.06 ± 10.46	61.71 ± 4.67	1830.45 ± 225.56	129.42 ± 6.98	79.53 ± 7.49	60.32 ± 8.63

FM Fat mass, *FFM* Fat free mass, *RMR* Resting metabolic rate, *SBP* Systolic blood pressure, *DBP* Diastolic blood pressure, *RHR* Resting heart rate. Data are shown as means ± SD

measured without shoes using a stadiometer (Leicester height measure, SECA; UK). Blood pressure and resting HR were measured using an automatic sphygmomanometer (Microlife BP A2 Basic, Gentle Technology, Microlife, Clearwater, FL, USA, Inc.). Next, body composition was estimated using a 2-compartment model via air displacement plethysmography (BodPod, Life Measurement, Inc.; USA). The Siri equation [30] was used to derive absolute and percentage fat mass (FM) and fat-free mass (FFM), while body weight was obtained from the BodPod scales. The BodPod has been demonstrated to show excellent accuracy for the estimation of body composition [31].

2.4 Resting metabolic rate

Resting metabolic rate (RMR) was measured in a dimly lit room, in the supine position for 30 min by an indirect calorimeter fitted with a ventilated hood (GEM, Nutren Technology Ltd.; UK). The GEM was calibrated in accordance with manufacturer's instructions prior to each measurement. Resting metabolic rate was calculated from VO_2 and VCO_2 in the steady state, defined as the 5 min block with the lowest coefficient of variation, after the removal of the first 5 min of data [32]. If RMR data were unavailable ($n = 2$), RMR was estimated a body mass index specific RMR algorithm of Müller [33].

2.5 Instruments

2.5.1 Polar HR monitor

HR was assessed during the PA protocol using a Polar m400 HR Monitor Watch (Polar Electro, Kempele, Finland) and a Polar H7 chest strap (Polar Electro, Kempele, Finland), which transmitted second-level data via a Bluetooth connection. Data were uploaded to the Polar flow online application, then downloaded and aggregated to minute-level for analysis. The Polar H7 served as a criterion measure of HR in the present study and it has been shown to have near perfect correlation with electrocardiogram during many exercise modalities [34].

2.5.2 Fitbit Charge 2

The FC2 (Fitbit Inc., San Francisco, CA, USA) is a wrist-worn activity monitor which estimates HR, steps, EE and PA, based on data obtained from incorporated sensors via proprietary algorithms. HR estimates are obtained through a patented technology called 'PurePulse', which uses light-emitting diodes on the surface of the skin to monitor blood volume continuously [35]. Data are aggregated to the minute-level and synced via the Fitbit mobile application to Fitbit servers through an application programming interface. Participants used the devices provided to them as part of the NoHoW trial

and if participants were not part of this trial a FC2 was provided for the duration of this study. The device was fitted a finger's width above the non-dominant wrist and was configured with participant weight, height, sex and date of birth.

2.5.3 SenseWear armband Mini

The SWA (BodyMedia Inc., Pittsburgh, PA) is a research-grade device which utilises a tri-axial accelerometer, heat-related sensors (heat flux, skin temperature, near body ambient temperature) and galvanic skin response to estimate EE. Data were downloaded and processed using the SenseWear® Pro 8.0 software, algorithm v5.2. The SWA was fitted with an elastic strap around the non-dominant arm and initialised using participant weight, height, sex, date of birth and smoking status.

2.5.4 Vyntus CPX

A stationary metabolic cart fitted with a respiratory facemask (Vyntus CPX, Jaeger-CareFusion, UK) was used as the criterion measure of EE in the present study. The Vyntus CPX has been demonstrated to be valid and to have excellent reliability (coefficient of variation <0.5%) [36] and is therefore used as a reference for the validation of portable systems [37]. The unit was calibrated prior to each lab visit in accordance with manufacturer's instructions. Breath by breath data from the device were aggregated to minute level and EE (kcal/min^{-1}) values were calculated from VO_2 and VCO_2 data assuming a minimal contribution of protein oxidation [38].

2.6 Statistical analysis

All analyses were conducted in R version 3.5.1 and Rstudio Version 1.1.447. Statistical significance was accepted at $p < 0.05$ for all analyses. Descriptive statistics (mean ± SD) were calculated for age, weight, height, FM, FFM and RMR. Data from the devices and criterion measures were averaged to provide mean HR in beats per minute (BPM) or EE (kcal/min^{-1}) for each participant. Data for each of the outputs were matched by time for each participant. Next, the first minute of data from each activity performed in the activity protocol was removed leaving minutes 2–5, which we considered as steady-state. These data were then averaged for each participant's activity bout and this figure was used in analyses.

Analyses for each of the devices, HR and EE were conducted separately. In line with previous research [39] we employed a range of statistical tests. Firstly, agreement between criterion measure and devices was assessed with Pearson's correlation coefficient. The method of Bland-Altman [40] was used to investigate mean difference between criterion and device estimates, with limits of agreement set to

± 1.96 x standard deviation of mean difference, using the 'BlandAltmanLeh' package in R. Root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE), were calculated with the R package 'metrics'. Lastly, equivalence tests were conducted to compare devices and criterion estimates using the 'TOSTpaired.raw' function within the 'TOSTER' package in R. For estimates to be considered equivalent, the 90% confidence interval needed to fall within the equivalence zone, which was considered to be ±10% of the criterion mean [41]. Lastly, the absolute percentage error, defined as the absolute value of the percentage error relative to the criterion were explored. Differences in absolute percentage error for sex were investigated with a one-way analysis of variance (ANOVA) and a post-hoc Tukey honest significant difference test, conducted using 'aov' from the 'stats' package in R. We investigated the relationship between continuous variables (age, RMR, height, weight, FM, FFM, resting HR, systolic and diastolic blood pressure) and absolute error rate in EE and HR estimates with Pearson's correlations, using the 'cor' function from the 'stats' package in R.

3 Results

The PA protocol was performed by all participants ($n = 59$) however the running task ($n = 49$), the 5% incline run ($n = 30$) and the moderate cycling tasks ($n = 58$) were not performed by all participants due to ranges in physical fitness within the sample.

3.1 Energy expenditure

3.1.1 Fitbit Charge 2

Synchronisation errors occurred for two participant's FC2 data and therefore 57 participant's data were included in FC2 analyses. The pooled result of all available bouts was a mean overestimation by the FC2 of 0.8 (kcal/min^{-1}), RMSE = 2.3 (kcal/min^{-1}), correlation coefficient of r = 0.77, MAPE = 44% and a non-significant equivalence test ($p > 0.05$) indicating that the FC2 was not equivalent to the criterion measure overall. The activity specific statistics, and the number of bouts included in the analyses are presented in Table 2. The poorest accuracy was observed in the folding and sweeping tasks, in which the FC2 overestimated with MAPE values of 93% and 81%, respectively (Fig. 1). The best accuracy, and statistical equivalence was observed in incline running tasks (MAPE = 12%). A Bland-Altman plot of the overall error is shown in Fig. 2, for which the 95% limits of agreement were: −3.52, 5.14 (kcal/min^{-1}).

Table 2 Statistics detailing the validity of EE estimates obtained from the FC2 (above) and SWA (below)

| Device	Activity	Bouts	ID	Device	Criterion	RMSE	MAPE	MAE	Correlation	Equivalence
FC2	sit	228	57	1.08 ± 0.24	1.30 ± 0.31	0.32	19	0.26	0.66	
	stand	228	57	1.15 ± 0.29	1.47 ± 0.35	0.44	24	0.37	0.56	
	walk	228	57	7.10 ± 1.97	4.27 ± 0.86	3.35	69	2.83	0.39	
	walk incline	228	57	7.32 ± 2.39	5.66 ± 1.02	2.56	31	1.75	0.59	
	run	191	48	9.91 ± 1.91	9.18 ± 1.83	1.61	15	1.29	0.70	
	run incline	120	30	10.61 ± 2.57	11.14 ± 2.22	1.58	12	1.27	0.81	Equivalent
	cycle low	225	57	3.78 ± 2.17	4.49 ± 1.23	2.15	40	1.70	0.38	
	cycle mid	217	56	4.35 ± 2.50	5.59 ± 1.54	2.69	39	2.14	0.37	
	folding	228	57	5.57 ± 1.88	2.96 ± 0.61	3.11	93	2.70	0.42	
	sweeping	228	57	5.98 ± 1.69	3.38 ± 0.83	2.94	81	2.64	0.58	
SWA	sit	236	59	1.43 ± 0.31	1.29 ± 0.31	0.25	17	0.20	0.75	
	stand	236	59	1.67 ± 0.36	1.47 ± 0.34	0.33	20	0.26	0.71	
	walk	236	59	4.47 ± 0.79	4.28 ± 0.85	0.73	14	0.59	0.62	Equivalent
	walk incline	236	59	5.12 ± 0.82	5.67 ± 1.00	1.02	13	0.78	0.56	
	run	195	49	9.73 ± 1.99	9.18 ± 1.81	1.60	15	1.34	0.69	
	run incline	120	30	9.69 ± 1.94	11.14 ± 2.22	2.14	15	1.76	0.71	
	cycle low	233	59	3.17 ± 1.19	4.51 ± 1.22	1.63	31	1.40	0.70	
	cycle mid	225	58	4.13 ± 1.98	5.60 ± 1.52	2.42	35	1.93	0.41	
	folding	236	59	5.31 ± 2.18	2.97 ± 0.60	3.06	83	2.43	0.43	
	sweeping	236	59	4.33 ± 1.70	3.37 ± 0.82	1.80	41	1.30	0.43	

Activity is laid out in the order dictated by the physical activity protocol. 'Bouts' refers to the number of activity bouts included and 'ID' refers to the number of participants included in each comparison. 'Correlation' refers to Pearson's R. 'Equivalence' refers to the results of the equivalence tests and the absence of text implies a non-significant equivalence test. Data are shown as means ± SD. MAPE = Mean absolute percentage error, RMSE = Root mean squared error, MAE = Mean absolute error

3.1.2 SenseWear Armband

EE data were available for all participants from the SWA and thus 59 participant's data were included in the SWA analyses. The pooled result of all available bouts was a mean overestimation of 0.03 (kcal/min^{-1}), RMSE = 1.7 (kcal/min^{-1}) correlation coefficient of r = 0.82, MAPE = 29% and a significant equivalence test ($p < 0.001$), indicating that the SWA was equivalent to the criterion measure overall. The activity specific statistics, and the number of bouts included in the

Fig. 1 A bar plot detailing the mean absolute percentage error (MAPE) of EE estimates from the SWA (yellow) and the FC2 (grey) for each of the activities performed in this study
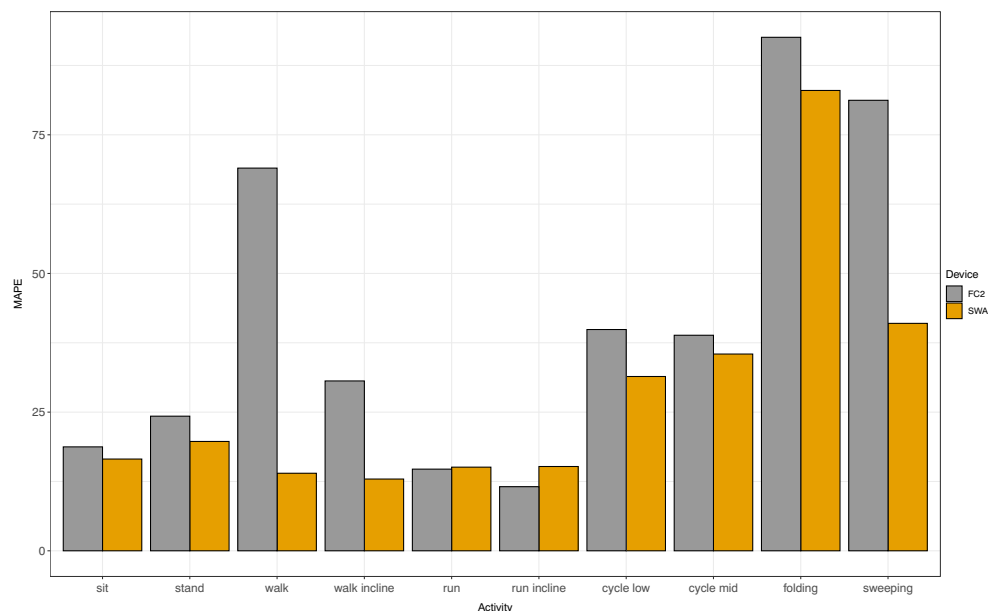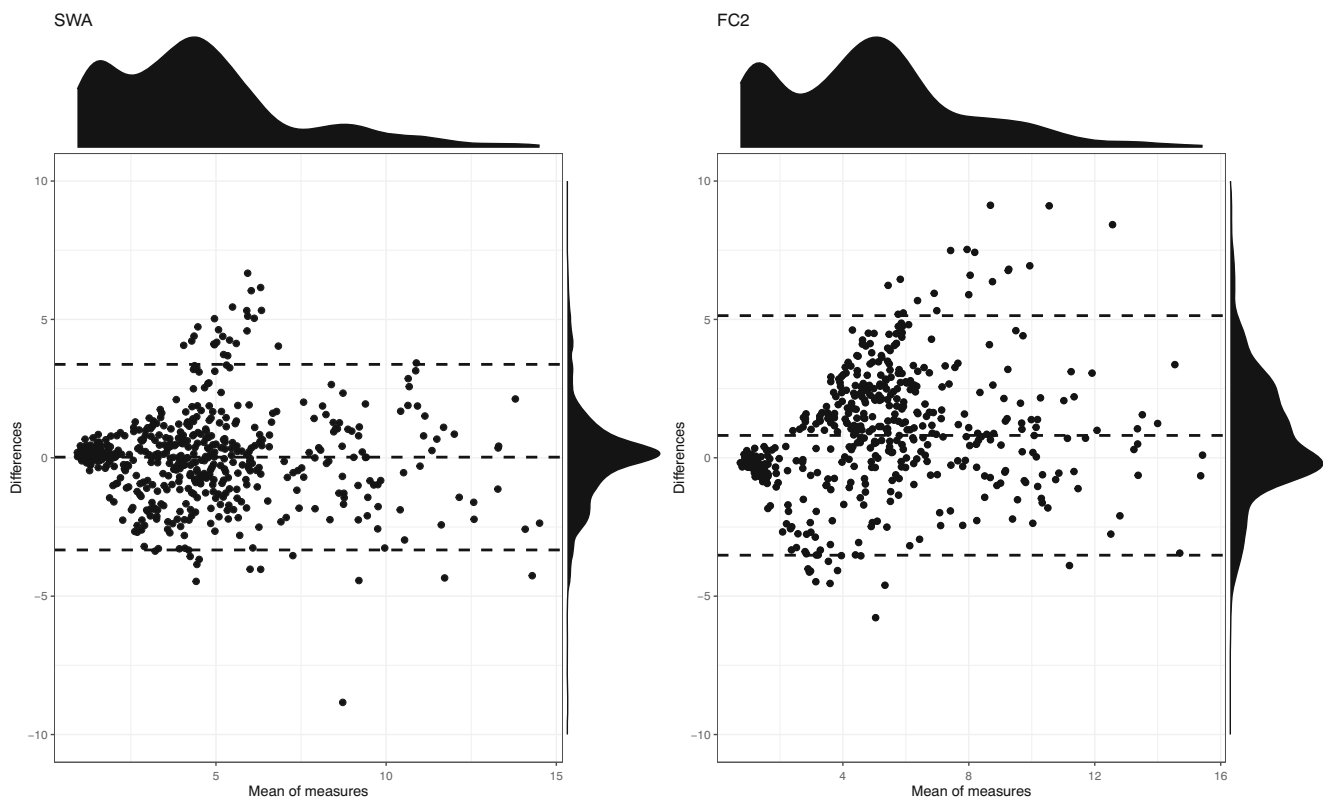
Fig. 2 Overall Bland-Altman plots of EE estimates from the SWA and FC2 relative to the criterion indirect calorimetry measure (Vyntus CPX). Data are displayed as kcal/min. 'Differences' represents device estimates – criterion estimates and is shown by the middle dashed line. The upper and lower dashed lines represent the upper and lower 95% limits. Mean of measures represents the average value of the criterion and device estimate. The density plots visualise the distribution of data points over the differences between the measures and the means of the measures

analyses are presented in Table 2. The SWA demonstrated the poorest accuracy in the folding task, in which it overestimated EE (MAPE = 83%). The lowest MAPE values were observed in the walking (MAPE = 14%) and walk 5% incline tasks (MAPE = 13%), which were overestimations and underestimations relative to the criterion measure, respectively (Fig. 1). Equivalence testing showed statistical equivalence between the SWA and the criterion measure during walking only. A Bland-Altman plot of the overall error is shown in Fig. 2, for which the 95% limits of agreement were: −3.33, 3.38 (kcal/min^{-1}).

3.1.3 Heart rate (HR)

Polar HR connectivity error occurred for one participant and thus HR analyses were conducted with 56 of the 57 participants with FC2 data. The pooled result of all available bouts was 98 ± 27 BPM (polar) vs 99 ± 29 BPM (FC2), RMSE = 20 BPM, correlation coefficient of r = 0.75, MAPE = 13% and a significant equivalence test ($p < 0.001$), indicating statistical equivalence. A Bland-Altman plot for errors in HR illustrates the agreement between criterion HR and FC2 HR by displaying the mean difference and 95% limits of agreement (Fig. 3) and the 95% limits of Agreement were: −37.94, 39.73 (BPM). Activity specific

Bland-Altman plots are presented for all tasks in Fig. 4 and accuracy statistics are presented in Table 3.

3.2 Predictors of absolute percentage error

Using the available data, no significant correlations were observed for any continuous variables and the absolute percentage error for HR and EE. ANOVA tests for the sex differences were not significant for EE absolute percentage errors for the SWA and FC2. In the HR comparison, a significant difference was observed between male bouts ($n = 184$) and female bouts ($n = 348$), with the absolute percentage error for males being significantly higher ($F = 4.158$, $p = 0.042$).

4 Discussion

This study investigated the validity of EE and HR estimates from the FC2 and EE estimates from the SWA in a heterogenous population performing a variety of tasks by comparing HR estimates to a HR chest strap (Polar) and EE estimates to a stationary metabolic cart (Vyntus CPX). The principal findings are i) the research-grade SWA was observed to be more accurate than the commercial-grade FC2 overall ii) the HR estimates of the FC2

Fig. 3 Overall Bland-Altman plots of HR estimates from the FC2 relative to the criterion measure (Polar chest strap). Data are displayed as beats per minute. 'Differences' represents device estimates – criterion estimates and is shown by the middle dashed line. The upper and lower dashed lines represent the upper and lower 95% limits. Mean of measures represents the average value of the criterion and device estimate. The density plots visualise the distribution of data points over the differences between the measures and the means of the measures
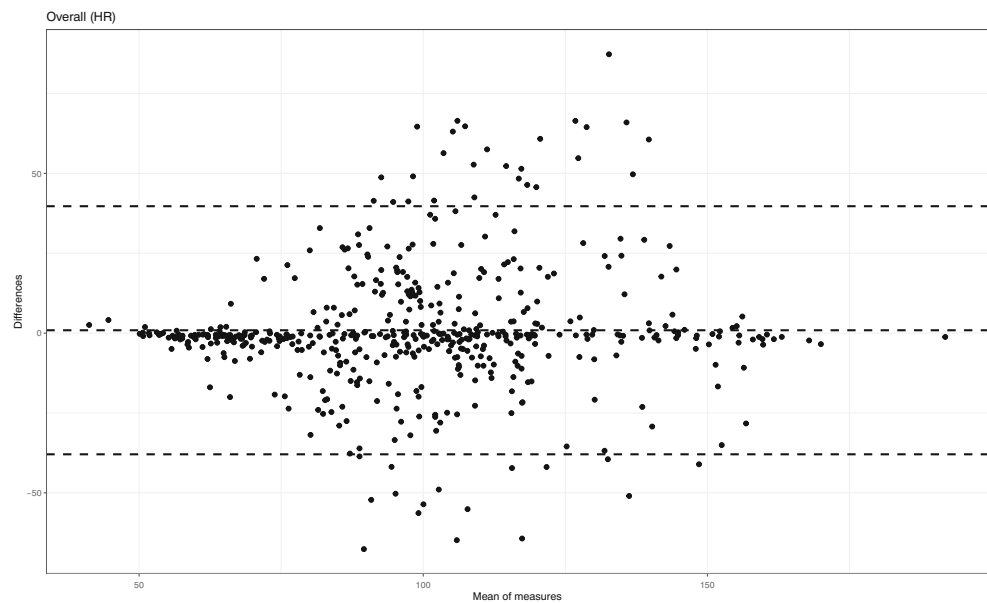
are generally in closer agreement with the criterion measures compared to EE estimates.

The FC2, one of the newest Fitbit activity monitors, has been investigated previously for its validity in estimating EE, relative to indirect calorimetry [19, 42], but this study provides a direct comparison with the SWA, a more established and commonly used research-grade device, using a range of activities. Our results substantiate previous research concluding that the SWA is more valid for the estimation of EE when compared to commercial activity monitors [21, 22]. This being said, the FC2 nor the SWA were consistently equivalent across the range of activities performed, with MAPE values >25% in some activities.

Large overestimations were observed for the FC2 during the household tasks. This most likely originates from the reliance on wrist accelerometry and this is a recognised limitation of devices located at this wear site [43]. Movements such as folding and sweeping, which involve rapid movements of the hand but are not particularly energetically demanding (typically ~4 metabolic equivalents) [44] were overestimated. This is opposite to the issue faced by more traditional devices, which were worn on the hip and underestimate the energy cost of tasks with limited ambulation (i.e. household tasks) [45, 46]. Notably, the MAPE values for the FC2 were lowest in running activities (indicating a high degree of accuracy) and higher during walking activities. This finding is reflective of the results of a recent meta-analysis published by our group, in which the pooled results from five comparisons for the Fitbit Charge HR (prior model to the FC2) showed significant, moderate to large overestimation relative to criterion measures of EE during ambulation and a non-significant overestimation during running [21]. Whilst we are limited in our ability to comment on the underlying cause of this error due to the proprietary nature of the algorithms, it is interesting to note that the greatest overestimate in HR estimates was observed in

the walking tasks. If HR is incorporated in the FC2 EE prediction algorithm, this could partially explain this result.

The performance of the SWA for the estimation of total daily EE is well recognised [47–49]. However, its accuracy in specific activity types is less established [50]. Indeed, significant underestimations relative to indirect calorimetry in running at higher speeds (> 9.9 km/h) have been reported [51] and in a validation study involving cycling, the SWA again significantly underestimated EE [52]. Data from the CALERIE study showed a mean bias in total daily EE estimates of -1.6 ± 261 kcal/d when compared with doubly labelled water, yet when the data were tertiled by total daily EE an underestimation of 162 kcal/d in the highest total daily EE group was observed [53]. The complimentary results overall and in comparisons to doubly labelled water may be largely influenced by the accuracy of the resting EE equations selected by the manufacturers, which are derived from participant characteristics [46]. The present results offer some support for this supposition and indicate that the SWA accuracy is dependent on the PA level of the individual.

The conclusion that the estimates of HR from the FC2 are typically more accurate than EE estimates is reflective of previous research [54, 55]. When HR estimates were aggregated across all available bouts, the HR estimates of the FC2 were statistically equivalent to the criterion measure. Error in specific activity types was greater but the FC2 was statistically equivalent in most activity types. A recent study reported that erratic movements and a greater HR were associated with an increased error in HR [56] and another concluded that the error was exacerbated with increasing exercise intensity [57]. In contrast, our results showed the highest error in the walking task, yet the greatest accuracy in the running and sedentary tasks. The observation of the greatest error in walking is similar to that reported in a previous study investigating the Fitbit Surge device
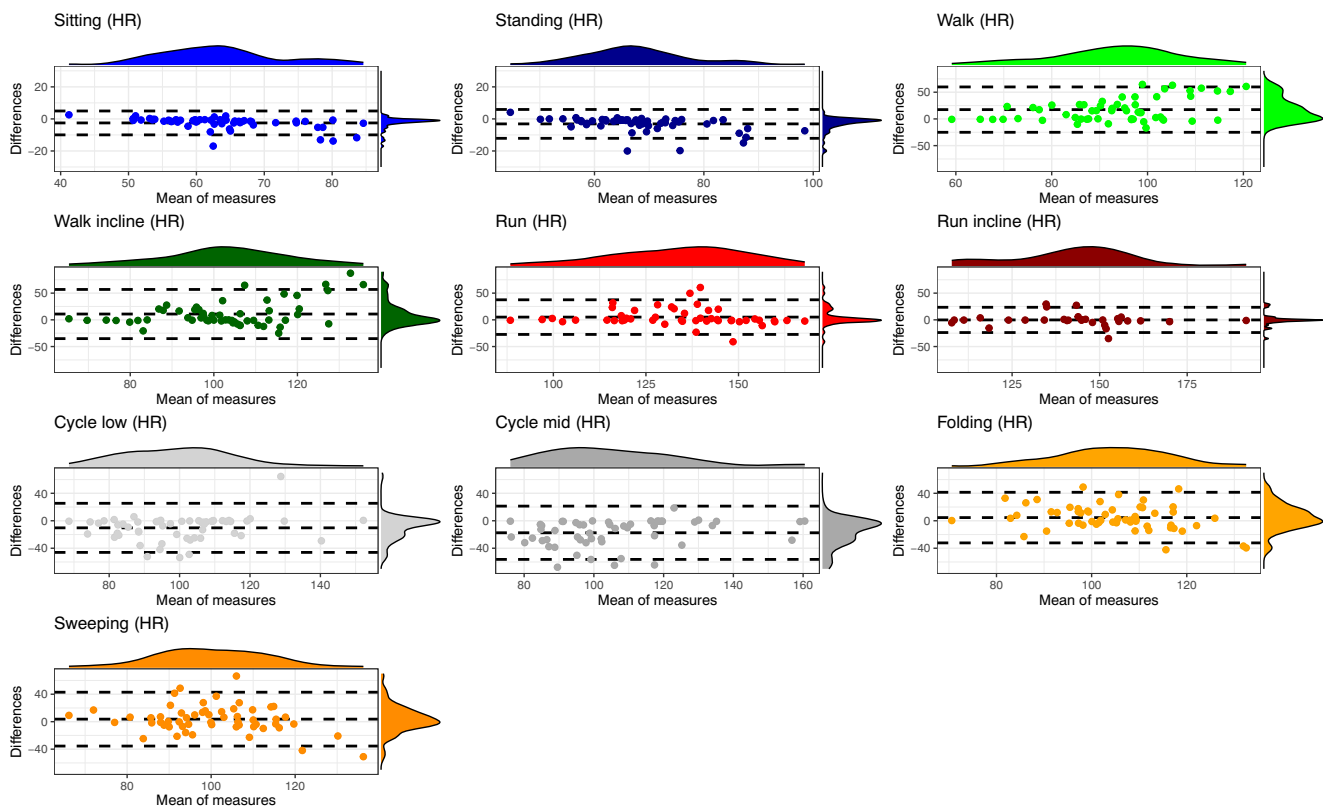
Fig. 4 Activity specific Bland-Altman plots for HR estimates from the FC2 relative to the criterion measure (Polar chest strap). Data are displayed as beats per minute. 'Differences' represents device estimates – criterion estimates and is shown by the middle dashed line. The upper and lower dashed lines represent the upper and lower 95% limits. Mean of measures represents the average value of the criterion and device estimate. The density plots visualise the distribution of data points over the differences between the measures and the means of the measures

which showed a greater error in HR during ambulatory tasks [55]. In contrast, two other studies investigating the FC2 report small underestimations in HR during walking [42, 56].

We identified no significant continuous correlates of the error for each device and this includes body composition, which we believe to be a novel investigation within this field. However, the percentage error in HR was significantly greater in males, when compared to females. Whilst the proprietary nature of the smoothing algorithms makes understanding the observed error challenging, photoplethysmography technology is likely to be influenced by device position and skin conditions which may differ between males and females [58]. Prior to the exercise condition the position and tightness of the FC2 were standardised for all participants and it therefore seems unlikely that this played a role in the observed error. It remains to be seen whether the free-living performance of the FC2 will differ between participants in less controlled environments and this should be addressed in future research.

4.1 Implications

The seeming inability of the 'out of the box' FC2 estimates to accurately estimate EE is a primary limitation for energy balance research, particularly when the numerous benefits of

cost, cloud storage and acceptance from participants are considered [59, 60]. Our data indicate that it may be more appropriate to use commercial activity trackers, in their current format, to infer PA from step counts or to estimate HR, which are generally observed to be more valid than EE estimates [17]. Alternatively, the application of metrics such as the heart rate reserve [26], which can be used to define minute level relative intensity from HR data may be preferred. These findings are important for studies utilising the FC2 for longitudinal data collection.

An accurate and objective estimate of EE, in combination with an estimate of change in energy storage, can be used to estimate energy intake [61] and therefore determine misreporting through the 'solving' of the energy balance equation [53]. Given the centrality of energy intake and EE to the development of obesity, it is vital to be able to estimate energy intake and EE with precision and accuracy in free-living individuals. Self-reported energy intake is still widely used in research, yet it is well established that this approach is limited by issues of misreporting [16]. Mathematical models to estimate energy intake from body weight have been developed and validated [62]. However, these models make assumptions about the EE levels, which are unlikely to be constant between and within individuals during weight loss and maintenance

Table 3 Statistics detailing the validity of HR estimates obtained from the FC2, measured in beats per minute

Activity	Bouts	ID	Device	Criterion	RMSE	MAPE	MAE	Correlation	Equivalence
sit	224	56	62.29 ± 8.38	64.80 ± 10.25	4.52	4	2.79	0.94	Equivalent
stand	224	56	66.44 ± 9.49	69.54 ± 11.54	5.51	4	3.31	0.92	Equivalent
walk	224	56	101.80 ± 20.59	84.40 ± 12.95	27.63	25	19.50	0.23	
walk incline	224	56	108.06 ± 22.94	97.19 ± 14.84	25.68	17	16.10	0.29	
run	191	48	136.15 ± 19.12	131.04 ± 20.93	17.16	8	10.02	0.66	Equivalent
run incline	120	30	142.13 ± 19.00	142.26 ± 20.15	11.85	5	6.81	0.81	Equivalent
cycle low	217	55	95.09 ± 20.55	105.40 ± 17.40	20.80	12	13.12	0.55	
cycle mid	209	54	97.29 ± 24.44	114.73 ± 19.68	26.25	16	18.17	0.62	
folding	224	56	106.67 ± 12.91	102.03 ± 17.94	19.20	15	14.38	0.29	Equivalent
sweeping	224	56	102.27 ± 14.76	98.55 ± 18.78	20.17	16	14.40	0.31	Equivalent

Activity is laid out in the order dictated by the physical activity protocol. 'Bouts' refers to the number of activity bouts included and 'ID' refers to the number of participants included in each comparison. 'Correlation' refers to Pearson's R. 'Equivalence' refers to the results of the equivalence tests and the absence of text implies a non-significant equivalence test. Data are shown as means ± SD. MAPE = Mean absolute percentage error, RMSE = Root mean squared error, MAE = Mean absolute error

interventions [1]. An inexpensive, objective estimate of EE will therefore improve energy intake estimates from mathematical models and whilst devices such as the FC2 show large inaccuracies, it is likely that in their current form, they would be superior than an estimation of constant PA EE.

Considering that it is possible to access minute-level data from commercial wearables in many instances, this raises the possibility of the application of non-linear modelling to improve estimates of EE from commercial wearable devices. Advanced statistical learning techniques are being used to estimate EE and PA of tasks with better accuracy than linear regression approaches [63–65] and future research should investigate whether data from commercial activity monitors can be used to more accurately predict EE from sensor outputs. The incorporation of body composition and participant characteristics to non-linear models could improve estimates of EE beyond the estimates of current activity monitors [66].

4.2 Limitations

In this study, a number of different FC2 devices were used and data were synced with each participant's mobile phone application. The lack of standardisation of devices may be considered a limitation, as different firmware could have been employed for different participants. However, this reflects the use of wearable devices in research environments, in which a study population are each provided with their own activity tracker and data are collected via an application programming interface.

Secondly, whilst this study provides analysis of the accuracy of two activity monitors for a relatively limited series of prescribed activities, it provides little insight into the ecological validity of these devices. Substantial over and underestimations from the FC2, depending on the specific activity in question, were observed and therefore the error observed in free-living individuals is likely to vary depending on the activities performed. Given that wearable devices will be used in free-living research, validation studies in free-living conditions are urgently required. Thirdly, this study was conducted in healthy, ambulatory individuals who were not pregnant, using medications associated with alteration to metabolic rate, and did not have cardiovascular, metabolic, renal disorders, illness or injury. It is possible that results would vary as the characteristics of study populations differ, however, with the exception gender difference in HR error, we found no evidence that this is the case.

5 Conclusion

The SWA is more valid for the estimation of EE when compared to the commercial grade FC2, yet neither activity monitor can consistently estimate EE with equivalence to a criterion measure. The FC2 provides better estimates of HR than it does EE, which are broadly, but not always, equivalent to criterion estimates across a broad range of activity types. It may therefore be more appropriate to focus on HR metrics for the assessment of PA, rather than EE in the FC2.

Compliance with ethical standards

Conflict of Interest None.

Informed consent Informed consent was obtained from all participants in this study.

References

1. Kerns JC, Guo J, Fothergill E, et al. Increased Physical Activity Associated with Less Weight Regain Six Years After "The Biggest Loser" Competition. Obesity. 2017;25:1838–43. https://doi.org/10.1002/oby.21986.

2. Wadden TA, Neiberg RH, Wing RR, et al. Four-year weight losses in the Look AHEAD study: factors associated with long-term success. Obesity (Silver Spring). 2011;19:1987–98. https://doi.org/10.1038/oby.2011.230.

3. Schoeller DA, Shay K, Kushner RF. How much physical activity is needed to minimize weight gain in previously obese women? Am J Clin Nutr. 1997;66:551–6. https://doi.org/10.1093/ajcn/66.3.551.

4. Hankinson AL, Daviglus ML, Bouchard C, et al. Maintaining a high physical activity level over 20 years and weight gain. JAMA - J Am Med Assoc. 2010. https://doi.org/10.1001/jama.2010.1843.

5. Swift DL, McGee JE, Earnest CP, et al. The Effects of Exercise and Physical Activity on Weight Loss and Maintenance. Prog Cardiovasc Dis. 2018. https://doi.org/10.1016/J.PCAD.2018.07.014.

6. MacLean PS, Wing RR, Davidson T, et al. NIH working group report: Innovative research to improve maintenance of weight loss. Obesity. 2015;23:7–15. https://doi.org/10.1002/oby.20967.

7. Beaulieu K, Hopkins M, Blundell J, et al. Impact of physical activity level and dietary fat content on passive overconsumption of energy in non-obese adults. Int J Behav Nutr Phys Act. 2017;14:14. https://doi.org/10.1186/s12966-017-0473-3.

8. Fruin ML, Rankin JW. Validity of a multi-sensor Armband in estimating rest and exercise energy expenditure. Med Sci Sports Exerc. 2004;36:1063–9. https://doi.org/10.1249/01.MSS.0000128144.91337.38.

9. Welk GJ, McClain JJ, Eisenmann JC, et al. Field Validation of the MTI Actigraph and BodyMedia Armband Monitor Using the IDEEA Monitor. Obes. 2007;15:918–28.

10. Lyden K, Kozey SL, Staudenmeyer JW, et al. A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations. Eur J Appl Physiol. 2011;111:187–201. https://doi.org/10.1007/s00421-010-1639-8.

11. Crouter SE, Clowers KG, Bassett DRJ. A novel method for using accelerometer data to predict energy expenditure. J Appl Physiol. 2006;100:1324–31. https://doi.org/10.1152/japplphysiol.00818.2005.

12. Brage S, Ekelund U, Brage N, et al. Hierarchy of individual calibration levels for heart rate and accelerometry to measure physical activity. J Appl Physiol. 2007;103:682–92. https://doi.org/10.1152/japplphysiol.00092.2006.

13. Bonomi AG, Plasqui G, Goris AHC, et al. Improving assessment of daily energy expenditure by identifying types of physical activity with a single accelerometer. J Appl Physiol. 2009;107:655–61. https://doi.org/10.1152/japplphysiol.00150.2009.

14. Black AE, Cole TJ. Within- and between-subject variation in energy expenditure measured by the doubly-labelled water technique: Implications for validating reported dietary energy intake. Eur J Clin Nutr. 2000;54:386–94. https://doi.org/10.1038/sj.ejcn.1600970.

15. Sardinha LB, Júdice PB. Usefulness of motion sensors to estimate energy expenditure in children and adults: A narrative review of studies using DLW. Eur J Clin Nutr. 2017;71:331–9. https://doi.org/10.1038/ejcn.2017.2.

16. Dhurandhar NV, Schoeller D, Brown AW, et al. Energy Balance Measurement: When Something is Not Better than Nothing. Int J Obes. 2015;39:1109–13. https://doi.org/10.1038/ijo.2014.199.

17. Feehan LM, Geldman J, Sayre EC, et al. Accuracy of fitbit devices: Systematic review and narrative syntheses of quantitative data. J Med Internet Res. 2018;20:e10527. https://doi.org/10.2196/10527.

18. Evenson KR, Goto MM, Furberg RD, et al. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Act. 2015;12:159. https://doi.org/10.1186/s12966-015-0314-1.

19. Boudreaux BD, Hebert EP, Hollander DB, et al. Validity of Wearable Activity Monitors during Cycling and Resistance Exercise. Med Sci Sports Exerc. 2018;50:624–33. https://doi.org/10.1249/MSS.0000000000001471.

20. Yang C-C, Hsu Y-L. A review of accelerometry-based wearable motion detectors for physical activity monitoring. Sensors (Basel). 2010;10:7772–88. https://doi.org/10.3390/s100807772.

21. O'Driscoll R, Turicchi J, Beaulieu K, et al. How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. Br J Sports Med. 2018;**77**:bjsports-2018-099643. https://doi.org/10.1136/bjsports-2018-099643.

22. Chowdhury EA, Western MJ, Nightingale TE, et al. Assessment of laboratory and daily energy expenditure estimates from consumer multisensor physical activity monitors. PLoS One. 2017;12:e0171720. https://doi.org/10.1371/journal.pone.0171720.

23. Spurr GB, Reina JC, Prentice a M, et al. Energy expenditure from minute-by-minute recording : comparison with indirect calorimetry. Am J Clin Nutr. 1988;48:552–9. https://doi.org/10.1093/ajcn/48.3.552.

24. Ceesay SM, Prentice AM, Day KC, et al. The use of heart rate monitoring in the estimation of energy expenditure : a validation study using indirect whole-body calorimetry. Brirish J Nutr. 1989;61:175–86. https://doi.org/10.1079/BJN19890107.

25. Karvonen MJ, Kentala E, Mustala O. The effects of training on heart rate; a longitudinal study. *Ann Med Exp Biol Fenn* 1957;**35**:307–15.http://www.ncbi.nlm.nih.gov/pubmed/13470504 (accessed 20 Oct 2018).

26. Schrack JA, Leroux A, Fleg JL, et al. Using Heart Rate and Accelerometry to Define Quantity and Intensity of Physical Activity in Older Adults. Journals Gerontol - Ser A Biol Sci Med Sci. 2018;73:668–75. https://doi.org/10.1093/gerona/gly029.

27. Achten J, Jeukendrup AE. Heart Rate Monitoring. Sports Med. 2003;33:517–38. https://doi.org/10.2165/00007256-200333070-00004.

28. Brage S, Westgate K, Franks PW, et al. Estimation of Free-Living Energy Expenditure by Heart Rate and Movement Sensing: A Doubly-Labelled Water Study. PLoS One. 2015;10:e0137206. https://doi.org/10.1371/journal.pone.0137206.

29. ACSM. Exercise Preparticipation Health Screen Recommendations. Published Online First: 2018. http://www.acsm.org/docs/default-source/publications/acsm-101-prescreeninginfographiccolorlegal-2015-12-15-v02.pdf?sfvrsn=2 (accessed 20 Feb 2018).

30. Siri WE. Body composition from fluid spaces and density: Analysis of methods. Adv Biol Med Phy. 1956.

31. Fields DA, Goran MI, McCrory MA. Body-composition assessment via air-displacement plethysmography in adults and children: A review. Am J Clin Nutr. 2002;75:453–67.

32. Sanchez-Delgado G, Alcantara JMA, Ortiz-Alvarez L, et al. Reliability of resting metabolic rate measurements in young adults:

Impact of methods for data analysis. Clin Nutr. 2018;37:1618–24. https://doi.org/10.1016/j.clnu.2017.07.026.

33. Müller MJ, Bosy-Westphal A, Klaus S, et al. World Health Organization equations have shortcomings for predicting resting energy expenditure in persons from a modern, affluent population: generation of a new reference standard from a retrospective analysis of a German database of resting energy expe. Am J Clin Nutr. 2004;80:1379–90. https://doi.org/10.1093/ajcn/80.5.1379.

34. Gillinov S, ETIWY M, Wang R, et al. Variable Accuracy of Wearable Heart Rate Monitors during Aerobic Exercise. Med Sci Sports Exerc. 2017;49:1697–703. https://doi.org/10.1249/MSS.0000000000001284.

35. Benedetto S, Caldato C, Bazzan E, et al. Assessment of the fitbit charge 2 for monitoring heart rate. PLoS One. 2018;13:e0192691. https://doi.org/10.1371/journal.pone.0192691.

36. Groepenhoff H, de Jeu RC, Schot R. Vyntus CPX compared to Oxycon pro shows equal gas-exchange and ventilation during exercise. In: *Respiratory Function Technologists/Scientists*. European Respiratory Society 2017. PA3002. doi: https://doi.org/10.1183/1393003.congress-2017.PA3002

37. Perez-Suarez I, Martin-Rincon M, Gonzalez-Henriquez JJ, et al. Accuracy and Precision of the COSMED K5 Portable Analyser. Front Physiol. 2018;9:1764. https://doi.org/10.3389/fphys.2018.01764.

38. Péronnet F, Massicotte D. Table of nonprotein respiratory quotient: an update. *Can J Sport Sci* 1991;**16**:23–9.http://www.ncbi.nlm.nih.gov/pubmed/1645211 (accessed 11 Jun 2019).

39. Bai Y, Hibbing P, Mantis C, et al. Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR. J Sports Sci. 2018;36:1734–41. https://doi.org/10.1080/02640414.2017.1412235.

40. Altman DG, Bland JM. Measurement in Medicine: the Analysis of Method Comparison Studies †. 1983. http://people.stat.sfu.ca/~raltman/stat300/AltmanBland.pdf (accessed 3 May 2019).

41. Lee J-MM, Kim Y-WY, Welk GJ. Validity of consumer-based physical activity monitors. Med Sci Sports Exerc. 2014;46:1840–8. https://doi.org/10.1249/MSS.0000000000000287.

42. Reddy RK, Pooni R, Zaharieva DP, 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:e10338. https://doi.org/10.2196/10338.

43. Ellis K, Kerr J, Godbole S, et al. Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification. Med Sci Sports Exerc. 2016;48:933–40. https://doi.org/10.1249/MSS.0000000000000840.

44. Ainsworth BE, Haskell WL, Herrmann SD, et al. 2011 Compendium of Physical Activities: a second update of codes and MET values. Med Sci Sports Exerc. 2011;43:1575–81. https://doi.org/10.1249/MSS.0b013e31821ece12.

45. Hendelman D, Miller K, Baggett C, *et al.* Validity of accelerometry for the assessment of moderate intensity physical activity in the field. *Med Sci Sports Exerc* 2000;**32**:S442–9. http://www.ncbi.nlm.nih.gov/pubmed/10993413 (accessed 3 Nov 2017).

46. Nelson MB, Kaminsky LA, Dickin DC, et al. Validity of Consumer-Based Physical Activity Monitors for Specific Activity Types. Med Sci Sports Exerc. 2016;48:1619–28. https://doi.org/10.1249/MSS.0000000000000933.

47. Slinde F, Bertz F, Winkvist A, et al. Energy expenditure by multisensor armband in overweight and obese lactating women validated by doubly labeled water. Obesity. 2013;21:2231–5. https://doi.org/10.1002/oby.20363.

48. Casiraghi F, Lertwattanarak R, Luzi L, et al. Energy Expenditure Evaluation in Humans and Non-Human Primates by SenseWear Armband. Validation of Energy Expenditure Evaluation by SenseWear Armband by Direct Comparison with Indirect

49. Johannsen DL, Calabro MA, Stewart J, et al. Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. Med Sci Sports Exerc. 2010;42:2134–40. https://doi.org/10.1249/MSS.0b013e3181e0b3ff.

50. Koehler K, Drenowatz C. Monitoring Energy Expenditure Using a Multi-Sensor Device-Applications and Limitations of the SenseWear Armband in Athletic Populations. Front Physiol. 2017;8:983. https://doi.org/10.3389/fphys.2017.00983.

51. Drenowatz C, Eisenmann JC. Validation of the SenseWear Armband at high intensity exercise. Eur J Appl Physiol. 2011;111:883–7. https://doi.org/10.1007/s00421-010-1695-0.

52. Koehler K, Braun H, de Marees M, et al. Assessing energy expenditure in male endurance athletes: Validity of the sensewear armband. Med Sci Sports Exerc. 2011;43:1328–33. https://doi.org/10.1249/MSS.0b013e31820750f5.

53. Shook RP, Hand GA, O'Connor DP, et al. Energy Intake Derived from an Energy Balance Equation, Validated Activity Monitors, and Dual X-Ray Absorptiometry Can Provide Acceptable Caloric Intake Data among Young Adults. J Nutr. 2018;148:490–6. https://doi.org/10.1093/jn/nxx029.

54. Wallen MP, Gomersall SR, Keating SE, et al. Accuracy of heart rate watches: Implications for weight management. PLoS One. 2016;11:e0154420. https://doi.org/10.1371/journal.pone.0154420.

55. Shcherbina A, Mattsson CM, Waggott D, et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. J Pers Med. 2017;7:1–12. https://doi.org/10.3390/jpm7020003.

56. Nelson BW, Allen N. Accuracy of Wearable Heart Rate During a Continuous and Ecologically Valid 24-Hour Period of Actual Consumer Device Use Conditions Within an Individual (Preprint). *JMIR mHealth uHealth* Published Online First. 2018. https://doi.org/10.2196/10828.

57. Thomson EA, Nuss K, Comstock A, et al. Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. J Sports Sci. 2019;37:1411–9. https://doi.org/10.1080/02640414.2018.1560644.

58. Stahl SE, An H-S, Dinkel DM, 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:e000106. https://doi.org/10.1136/bmjsem-2015-000106.

59. Wright SP, Hall Brown TS, Collier SR, et al. How consumer physical activity monitors could transform human physiology research. Am J Physiol Integr Comp Physiol. 2017;312:R358–67. https://doi.org/10.1152/ajpregu.00349.2016.

60. Gualtieri L, Rosenbluth S, Phillips J. Can a Free Wearable Activity Tracker Change Behavior? The Impact of Trackers on Adults in a Physician-Led Wellness Group. *JMIR Res Protoc*. 2016;5:e237. https://doi.org/10.2196/resprot.6534.

61. Racette SB, Das SK, Bhapkar M, et al. Approaches for quantifying energy intake and %calorie restriction during calorie restriction interventions in humans: the multicenter CALERIE study. AJP Endocrinol Metab. 2012;302:E441–8. https://doi.org/10.1152/ajpendo.00290.2011.

62. Sanghvi A, Redman LM, Martin CK, et al. Validation of an inexpensive and accurate mathematical method to measure long-term changes in free-living energy intake. Am J Clin Nutr. 2015;102:353–8. https://doi.org/10.3945/ajcn.115.111070.

63. Ellis K, Kerr J, Godbole S, et al. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol Meas. 2014;35:2191–203. https://doi.org/10.1088/0967-3334/35/11/2191.

64. Montoye AHKK, Conger SA, Connolly CP, et al. Validation of Accelerometer-Based Energy Expenditure Prediction Models in Structured and Simulated Free-Living Settings. Meas Phys Educ

Exerc Sci. 2017;21:1–12. https://doi.org/10.1080/1091367X.2017.1337638.

65. Staudenmayer J, Pober D, Crouter S, et al. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. J Appl Physiol. 2009;107:1300–7. https://doi.org/10.1152/japplphysiol.00465.2009.

66. Weyer C, Snitker S, Rising R, et al. Determinants of energy expenditure and fuel utilization in man: effects of body composition, age, sex, ethnicity and glucose tolerance in 916 subjects. Int J Obes. 1999;23:715–22. https://doi.org/10.1038/sj.ijo.0800910.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.