technologies

MDPI

Article

Determining the Reliability of Several Consumer-Based Physical Activity Monitors

Joshua M. Bock [1,2], Leonard A. Kaminsky [1,3], Matthew P. Harber [1] and Alexander H. K. Montoye [1,4,*]

1 Clinical Exercise Physiology Program, Ball State University, Muncie, IN 47306, USA; jmbock@healthcare.uiowa.edu (J.M.B.); kaminskyla@bsu.edu (L.A.K.); mharber@bsu.edu (M.P.H.)
2 Department of Physical Therapy and Rehabilitation Science, University of Iowa, Iowa City, IA 52242, USA
3 Fisher Institute of Health and Well-Being, Ball State University, Muncie, IN 47306, USA
4 Department of Integrative Physiology and Health Science, Alma College, Alma, MI 48801, USA
* Correspondence: montoyeah@alma.edu; Tel.: +1-989-463-7923

Received: 31 May 2017; Accepted: 21 July 2017; Published: 24 July 2017

Abstract: Limited research exists on the reliability of consumer-based physical activity monitors (CPAMs) despite numerous studies on their validity. Consumers often purchase CPAMs to assess their physical activity (PA) habits over time, emphasizing CPAM reliability more so than their validity; therefore, the purpose of this study was to investigate the reliability of several CPAMs. In this study, 30 participants wore a pair of four CPAM models (Fitbit One, Zip, Flex, and Jawbone Up24) for a total of eight monitors, while completing seven activities in the laboratory. Activities were completed in two consecutive five-minute bouts. Participants then wore either all wrist- or hip-mounted CPAMs in a free-living setting for the remainder of the day. Intra-monitor reliability for steps (0.88–0.99) was higher than kcals (0.77–0.94), and was higher for hip-worn CPAMs than for wrist-worn CPAMs ($p < 0.001$ for both). Inter-monitor reliability in the laboratory for steps (0.81–0.99) was higher than kcals (0.64–0.91) and higher for hip-worn CPAMs than for wrist-worn CPAMs ($p < 0.001$ for both). Free-living correlations were 0.61–0.98, 0.35–0.96, and 0.97–0.98 for steps, kcals, and active minutes, respectively. These findings illustrate that all CPAMs assessed yield reliable estimations of PA. Additionally, all CPAMs tested can provide reliable estimations of physical activity within the laboratory but appear less reliable in a free-living setting.

Keywords: physical activity; accelerometry; steps; energy expenditure; activity tracker

1. Introduction

Despite the well-known benefits of regularly engaging in physical activity (PA), half, or more, of U.S. adults do not meet the 2008 Physical Activity Guidelines for Americans [1–3]. To better understand the role of PA in improving health and reducing disease burden, it is important to measure PA accurately and reliably. High-quality measurement techniques allow researchers to identify which activity intensities provide optimal health benefits, monitor intermittent bouts of PA, and more accurately assess the effectiveness of interventions for promoting behavior change [4]. Consumer-based PA monitors (CPAMs) are common accessories with one in ten adults in the United States owning a CPAM [5]. During the first fiscal quarter of 2016, 19.7 million fitness trackers were sold worldwide; a 67.2% increase from quarter one of 2015. Fitbit Inc. was the largest distributor of fitness trackers during quarter one of 2015 and 2016 with 32.6 and 24.5% market share, respectively [6]. Despite surging popularity of these devices, one in three consumers who purchase a CPAM stops using it after six months [5]. The reasons for the high dropout in using CPAMs are not well understood, but they may be partly related to a lack of understanding on how well (e.g., accurately and reliably) the CPAMs capture PA levels and patterns over time.

While studies of device accuracy are common, much less research has investigated CPAM reliability [7]. For instance, only one study has assessed intra-monitor reliability (e.g., test-retest reliability). Kooiman et al. [8] assessed the intra-monitor reliability of the Fitbit Flex (Fitbit Inc., San Francisco, CA), Jawbone Up (AliphCom dba Jawbone, San Francisco, CA, USA), and Fitbit Zip (Fitbit Inc., San Francisco, CA, USA) to estimate steps using two bouts on a treadmill at 3.0 mph for 30 min. High intra-class correlations (ICCs) were found (0.81–0.90, for the Fitbit Flex, Fitbit Zip, and Jawbone Up), but these results are limited to a single activity and did not assess other variables, such as kcals or active minutes [8].

Four studies have assessed inter-monitor reliability (agreement between various monitors used during the same assessment). These studies ranged from case studies to those with 30 participants and assessed the Fitbit Ultra, Fitbit One, and Fitbit Flex. All four studies found Pearson correlations >0.90 for both steps and kcal measurements during ambulation in laboratory settings [9–11] or across free-living settings [12].

Current CPAM reliability research is limited regarding the diversity of activities tested (mainly walking and jogging) and the variables assessed (mainly steps). Additionally, several studies have evaluated inter-monitor reliability exclusively despite intra-monitor reliability being more relevant to assessing PA habits over time as consumers rarely use multiple CPAMs at a given time. Furthermore, little work has been done to assess CPAM inter-monitor reliability in a free-living setting. The inclusion of multiple settings is critical as several studies have reported setting-oriented differences in CPAM performance [13,14]. This study's purpose was to assess the intra- and inter-monitor reliability of several CPAMs for steps and kcals during a variety of activities, as well as the inter-monitor reliability to estimate steps, kcals, and active minutes in a free-living setting.

2. Methods

Participants

In this study, 30 (9M/21F) young adults were recruited from the East-Central region of Indiana. To be eligible for this study, participants had to be free of gait abnormalities, free of acute illness, between the ages of 18 and 80 years, not pregnant, and capable of completing the protocol without undue fatigue.

Prior to participating in the study all participants provided written informed consent approved by Ball State University's Institutional Review Board. All participants were right-handed and Caucasian; demographic information is shown in Table 1.

Table 1. Demographic information on participants categorized per analysis.

	All Participants	ICCs (n = 28)	Pearson (n = 30)	FL Hip (n = 15)	FL Wrist (n = 15)
Age (years)	23.1 ± 2.1	23.0 ± 2.1	23.0 ± 2.0	23.8 ± 2.4	22.4 ± 1.7
BMI ($kg \cdot m^{-2}$)	23.3 ± 3.4	23.4 ± 3.5	23.2 ± 3.3	23.3 ± 2.7	23.3 ± 4.0
Treadmill Brisk ($km \cdot h^{-1}$)	5.3 ± 0.3	5.3 ± 0.3	5.5 ± 0.3	- -	- -
Treadmill Jog ($km \cdot h^{-1}$)	8.7 ± 2.1	8.9 ± 2.1	8.9 ± 2.1	- -	- -

Kcal = kilocalories. BMI = body mass index. ICCs = data from participants used during intra-monitor analysis. Pearson = data from participants used during inter-monitor analysis. FL Hip = data from participants assigned hip-worn monitors during free-living portion of study. FL Wrist = data from participants assigned wrist-worn monitors during free-living portion of study. Data presented as mean ± standard deviation.

3. Equipments

During the laboratory visit, participants wore eight CPAMs (one pair of four different models). Descriptions of the CPAMs used are provided below.

Fitbit One (FO; Fitbit Inc., San Francisco, CA, USA): The FO, a hip-worn CPAM weighing 8.5 grams was used to estimate steps and kcals in the laboratory setting, as well as steps, kcals, and active minutes

during the free-living portion of the study. Data are quantified by the FO by utilizing the demographic information entered into the monitor, as well as through measurements made via accelerometer hardware within the monitor. This CPAM has an internal, rechargeable battery and provides real-time feedback to its user. The FO has the capability to synchronize with the Fitbit Mobile Application via a Bluetooth connection allowing the user to track PA over time. Data from the FO were collected from the Fitbit Mobile Application before and after each activity.

Fitbit Zip (Fitbit Inc., San Francisco, CA, USA): The FZ, is a hip-worn CPAM weighing 8.5 grams and was used to estimate steps and kcals in the laboratory setting as well as steps, kcals, and active minutes during the free-living portion of the study. Data are quantified by the FZ by utilizing the demographical information entered into the monitor, as well as through measurements made via the accelerometer hardware within the monitor. The FZ uses a CR-2032 watch battery and has the capability to synchronize with the Fitbit Mobile Application via a Bluetooth connection. Data from the FZ were collected from the device's built-in display screen before and after each activity.

Jawbone Up24 (AliphCom dba Jawbone, San Francisco, CA, USA): The JU, a wrist-worn CPAM weighing 22.7 grams and was used to estimate steps and kcals in the laboratory setting, as well as steps, kcals, and active minutes during the free-living portion of the study. Data are quantified by the JU by utilizing the demographical information entered into the monitor, as well as through measurements made via accelerometer hardware within the monitor. This CPAM utilizes an internal, rechargeable battery and can provide real-time feedback to its user via Bluetooth connection and the UP Mobile Application. Data from the JU were collected from the UP Mobile Application before and after each activity.

Fitbit Flex (Fitbit Inc., San Francisco, CA, USA): The FF, a wrist-worn CPAM weighing 17.0 grams and was used to estimate steps and kcals in the laboratory setting, as well as steps, kcals, and active minutes during the free-living portion of the study. Data are quantified by the FF by utilizing the demographical information entered into the monitor, as well as through measurements made via accelerometer hardware within the monitor. This monitor utilizes an internalized, rechargeable battery and requires the Fitbit Mobile Application and a Bluetooth connection to track PA. Data from the FF were collected from the Fitbit Mobile Application before and after each activity.

4. Protocol

Participants came to the Clinical Exercise Physiology Laboratory at Ball State University twice. During visit 1, participants completed an informed consent and had their height and weight measurements taken via scale (to the nearest 0.1 kg) and stadiometer (to the nearest 1.0 cm), which were then entered into each CPAM's respective mobile applications in addition to age, sex, and hand dominance. Researchers then fitted the CPAMs to the participants; initial readings of steps and kcals were collected from all CPAMs while the participant was in a seated position.

Following baseline data collection, participants completed a laboratory-based activity protocol. Each participant underwent an identical protocol where all activities lasted for five minutes, excluding transition time between activities. The only exception was the 'climbing stairs' activity in which all participants ascended and descended a flight of stairs five times at a self-selected pace. All activities were performed twice in succession with CPAM data collected before and after each activity bout to allow for the intra-monitor reliability analysis. It should be noted that data collected from the CPAMs were done so in the same order (FO, FZ, FF, JU) to minimize variability. Additionally, transition times between activity bouts lasted approximately one to three minutes and were determined by CPAM synchronization rate following activity bouts. The activity protocol was structured in the following order: typing, reading, sweeping (participants swept confetti into a pile within a ~10 m^2 section of the laboratory), slow treadmill walk at 3.2 km/h, brisk treadmill walk (4.8–5.6 km/h), treadmill jog (6.4–12.9 km/h), and ascending/descending stairs. Participants chose paces for the brisk treadmill walk, treadmill jog, and stairs activities.

5. Data Cleaning and Analysis

Intra-monitor reliability was assessed via intra-class correlations (ICCs) independently for each CPAM model (FO, FZ, JU, and FF) and outcome variable of interest (steps and kcals). Data used in this analysis came from a single monitor of each brand during both bouts of each activity. For wrist-worn CPAMs, data from the distal monitor were used, whereas data for hip-worn CPAMs came from the anterior monitor. Inter-monitor reliability was assessed by comparing data from each monitor brand with its pair (e.g., one FF against the other FF) for the first activity bout exclusively. Pearson correlations, calculated for each CPAM model and outcome variable of interest, were used to define the inter-monitor reliability for each CPAM. Both intra- and inter-monitor reliability analyses used protocol-wide data. That is, for each participant, there was a single ICC and Pearson correlation calculated using data from all activities, for a total of 30 data points for each analysis points. It should be noted that these analyses occurred after exclusion criteria were applied (see below).

Participants also completed a free-living protocol after their first laboratory visit. During this protocol, the participants continued to wear either hip-worn (FOs and FZs) or wrist-worn (FFs and JUs) CPAMs. Participants were assigned either hip- or wrist-worn CPAMs as the researchers presumed wearing all eight CPAMs for most of a day would be uncomfortable for participants and may, therefore, alter their behavior and/or reduce compliance with wearing the devices. These CPAMs were worn for the remainder of the day then returned to the lab the following morning (visit two) when the research staff collected CPAM monitors and data concluding participants' involvement in the study. The CPAMs assigned to the participants were arranged among participants so that each placement site (hip or wrist) was used by 15 participants. Free-living data were analyzed using Pearson correlations in a similar fashion to the laboratory inter-monitor reliability analysis.

A pair of exclusion criteria was applied to the collected CPAM data to remove data likely influenced by monitor malfunctions. The exclusion criteria for laboratory data were (1) data were negative (e.g., steps decreased following an activity) or (2) the kcals variable was not updated for a given CPAM following an activity. Exclusion criteria for the free-living portion were (1) data were negative or (2) steps taken over the remainder of the day were $\leq$150 steps. Once these criteria were applied, a repeated-measures analysis of variance (RM-ANOVA) with Tukey's post-hoc was used to determine if significant differences existed among the ICCs. Bland-Altman plots were created using step and kcal data from both bouts (intra) and monitor pairs (inter) to illustrate the nature of CPAM differences.

Additionally, median absolute differences (MAD) have been used alongside correlations to characterize agreement, as done in previous work [15]. Initially, absolute differences were calculated for each monitor per participant and activity. Then, the medians of the absolute differences were determined per monitor and participant and presented as MAD. Median percent difference (MPD) was calculated in a similar fashion using percent differences in place of absolute differences. MAD and MPD were calculated using step and kcal data from both bouts (intra-monitor) and each pair of monitors (inter-monitor).

All analyses were conducted in SPSS version 23.0 (IBM, Armonk, NY, USA) and Microsoft Excel (Microsoft, Redmond, WA). Statistical significance was defined a priori as $\alpha < 0.05$. Nomenclature for correlation strength was designated as follow: high (r = 0.80–1.00), moderately high (r = 0.60–0.79), low (r = 0.40–0.59), or no relationship (r = 0.00–0.19) as set forth by Safrit et al. [16].

6. Results

6.1. Intra-Monitor Reliability

Two participants were excluded from the intra-monitor reliability analysis due to errors encountered during data collection (e.g., poor synchronization of the mobile application), resulting in 28 participants' data being used during analysis. Additionally, exclusion criteria removed 11.8% of step and 8.3% of kcal data from participants included in the analysis. The ICCs for steps and kcals are

shown in Figure 1. All step ICCs were high (≥0.80) with the FZ significantly higher than the JU and FF ($p < 0.05$). All ICCs for kcals were moderately high (0.60–0.79) with the FO having significantly higher reliability than the JU and FF ($p < 0.05$). Data for MAD and MPD are shown in Table 2. Recorded estimations per CPAM and activity from the first visit are shown in Table 3. Intra-monitor MAD and MPD step values for the hip-worn CPAMs were significantly lower (better) than the wrist-worn CPAMs ($p < 0.01$) without a concurrent difference in kcal estimations ($p = 0.46$ and 0.53, respectively). The JU had the largest average MAD for steps (11), kcals (2.1), and the highest MPD for kcals (13.9%); the FF had the largest MPD for steps (7.2%). Figure 2 (steps) and Figure 3 (kcals) illustrate that CPAM error was higher in some cases during activities with higher predicted PA; however, these results may be partly influenced by outliers as they were not excluded from analysis. In general, the 95% limits of agreement were narrower for the hip-worn CPAMs compared to the wrist-worn CPAMs.

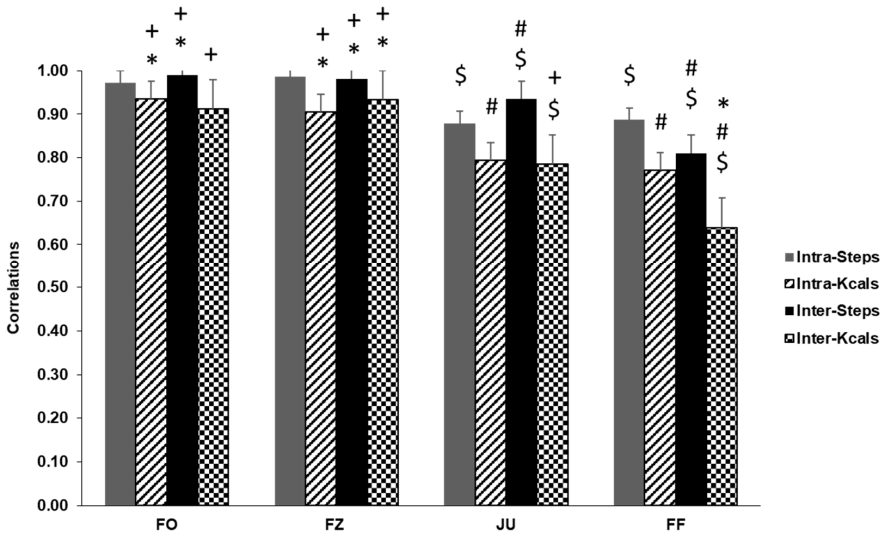

Figure 1. Intra-class correlations (Intra-) and Pearson correlations (inter-) for laboratory data. FO = Fitbit One. FZ = Fitbit Zip. JU = Jawbone Up24. FF = Fitbit Flex. Intra-Steps = intra-class coefficient for steps. Intra-kcals = intra-class coefficient for Calories. Inter-Steps = Pearson correlations for steps. Inter-Kcals = Pearson correlation for Calories. # statistically different from FO. $ statistically different from FZ. * statistically different from JU. + statistically different from FF. Statistical significance was defined as $p < 0.05$ for all.

Table 2. Median absolute differences and median absolute percent differences across entire laboratory protocol.

	FO	FZ	JU	FF
Intra-monitor reliability				
Steps	1.9 + (0.4) *	3.3 + (0.7) *	7.5 (2.3) #,$	10.9 #,$ (2.6)
Kcals	1.5 (8.8)	1.5 (9.1)	1.8 (12.5)	1.5 (8.9)
Inter-monitor reliability (Lab)				
Steps	0.5 *,+ (0.1) *,+	0.5 (0.1) +	2.8 # (0.7) #	4.0 # (1.4) #,$
Kcals	0.5 *,+ (4.6) *,+	0.5 *,+ (5.9) *,+	1.3 #,$ (9.4) #,$	1.5 #,$ (9.1) #,$
Inter-monitor reliability (FL)				
Steps	35 (2.1)	128 (7.0)	731 (8.1)	154 (5.2)
Kcals	34 (5.1)	60 (8.3)	26 (4.1)	88 (11.5)
Active Minutes	0 (0.0)	0 (0.0)	6 (8.5)	0 (0.0)

Kcals = kilocalories. Data presented as MAD (MPD). MAD = median absolute difference. MPD = median absolute percent difference. FL = free-living. FO = Fitbit One. FZ = Fitbit Zip. JU = Jawbone Up24. FF = Fitbit Flex. # significantly different from FO ($p < 0.05$). $ significantly different from FZ. * significantly different from JU. + significantly different from FF.

Table 3. Physical activity estimations per monitor and activity from visit one.

Steps	Typing	Reading	Sweeping	Slow TM	Brisk TM	TM Jog	Stairs
FO	0 ± 1	0 ± 0	10 ± 43	476 ± 35	587 ± 31	725 ± 105	113 ± 21
FZ	2 ± 8	1 ± 3	1 ± 3	461 ± 60	583 ± 100	744 ± 105	123 ± 36
JU	2 ± 6	0 ± 2	266 ± 166	426 ± 74	575 ± 127	779 ± 136	125 ± 46
FF	1 ± 3	3 ± 8	327 ± 116	399 ± 142	529 ± 151	757 ± 185	153 ± 128
Kcals							
FO	9.2 ± 2.0	7.7 ± 1.4	10.1 ± 2.6	20.8 ± 4.4	32.7 ± 6.4	48.9 ± 12.3	13.2 ± 3.7
FZ	11.6 ± 8.7	7.7 ± 1.2	7.8 ± 2.0	41.3 ± 25.8	41.3 ± 6.7	56.0 ± 9.0	12.1 ± 4.4
JU	10.7 ± 3.9	7.6 ± 1.7	17.6 ± 7.1	21.5 ± 7.4	31.1 ± 9.6	56.2 ± 22.4	12.1 ± 11.3
FF	12.8 ± 15.6	8.1 ± 1.7	28.9 ± 9.5	33.0 ± 11.7	39.1 ± 11.9	58.7 ± 18.3	14.5 ± 7.8

TM = treadmill. Kcal = kilocalories. FO = Fitbit One. FZ = Fitbit Zip. JU = Jawbone Up24. FF = Fitbit Flex. Data presented as mean ± standard deviation.

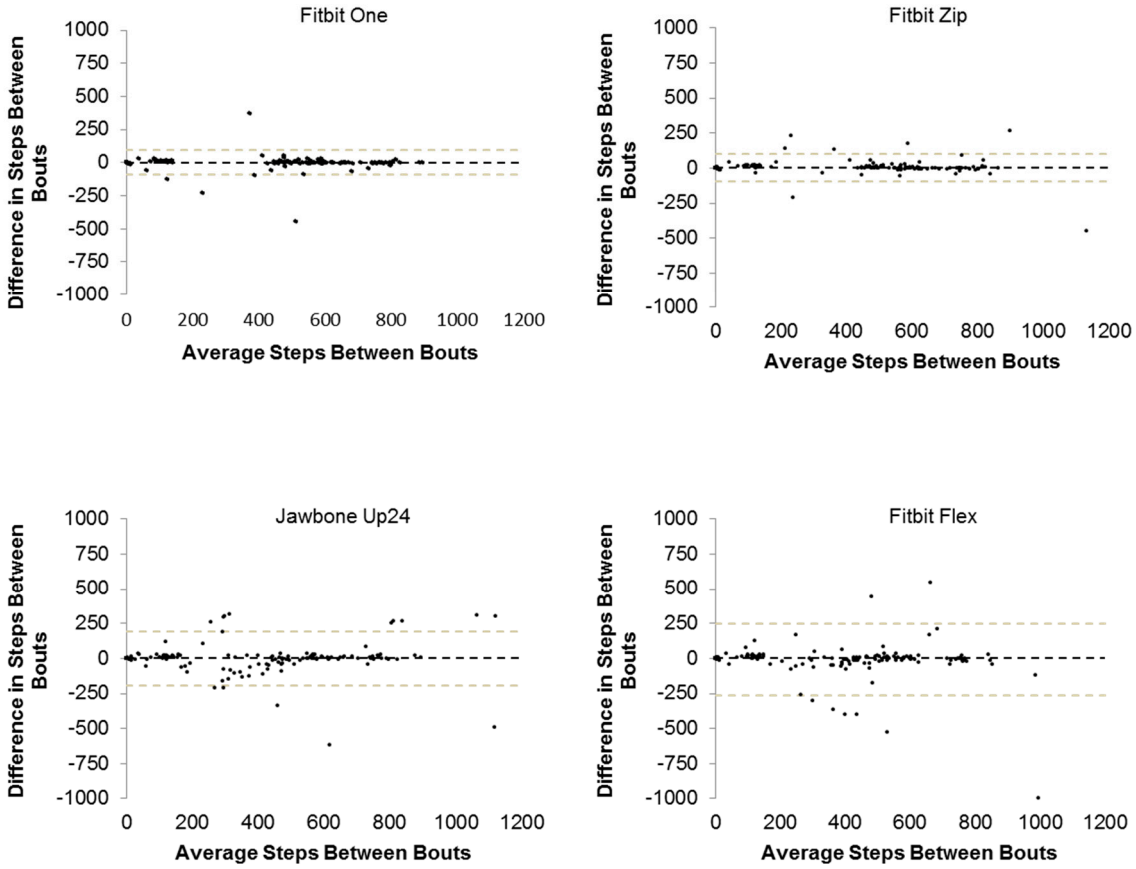

Figure 2. Bland-Altman plots with 95% limits of agreement calculated using the intra-monitor step data from all activities completed by each participant.

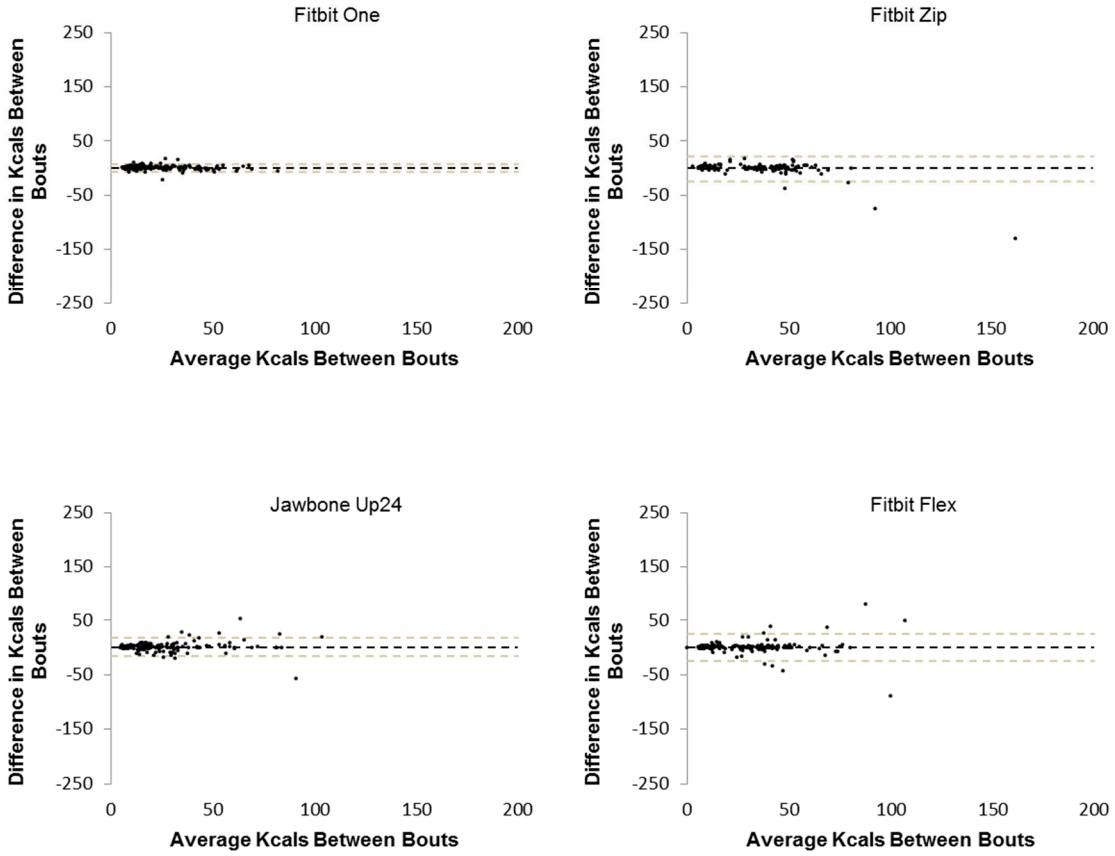

Figure 3. Bland-Altman plots with 95% limits of agreement calculated using the intra-monitor kilocalorie (kcals) data from all activities completed by each participant.

6.2. Inter-Monitor Reliability: Laboratory Setting

All 30 participants' data were included in the inter-monitor reliability analysis. Correlations for steps and kcals are shown in Figure 1. Prior to analysis, 11.5% (step) and 8.2% (kcal) data were removed per exclusion criteria mentioned above. All step correlations were high (≥0.80). Both hip-worn CPAMs (FO and FZ) had correlations significantly higher than the wrist-worn CPAMs (JU and FF, $p < 0.05$). Kcal correlations for the FO and FZ were high (≥0.80); the JU and FF correlations were moderately high (0.60–0.79). Correlations were significantly higher for the FO than the FF, the FZ than the JU and FF, and the JU than the FF ($p < 0.05$). Results from MAD and MPD are shown in Table 2. Recorded estimations per CPAM and activity from the first visit are shown in Table 3. Inter-MAD and MPD values were significantly lower in hip-worn CPAMs than wrist-worn CPAMs ($p < 0.05$ and < 0.01, respectively). The JU had the largest MAD and MPD for kcals (2.7 and 14.2%), the FF had the largest MAD value for steps (7), and the FZ had the largest MPD for steps (6.2%). For both steps (Figure 4) and kcals (Figure 5), the 95% limits of agreement were narrower for the hip-worn CPAMs compared to the wrist-worn CPAMs.

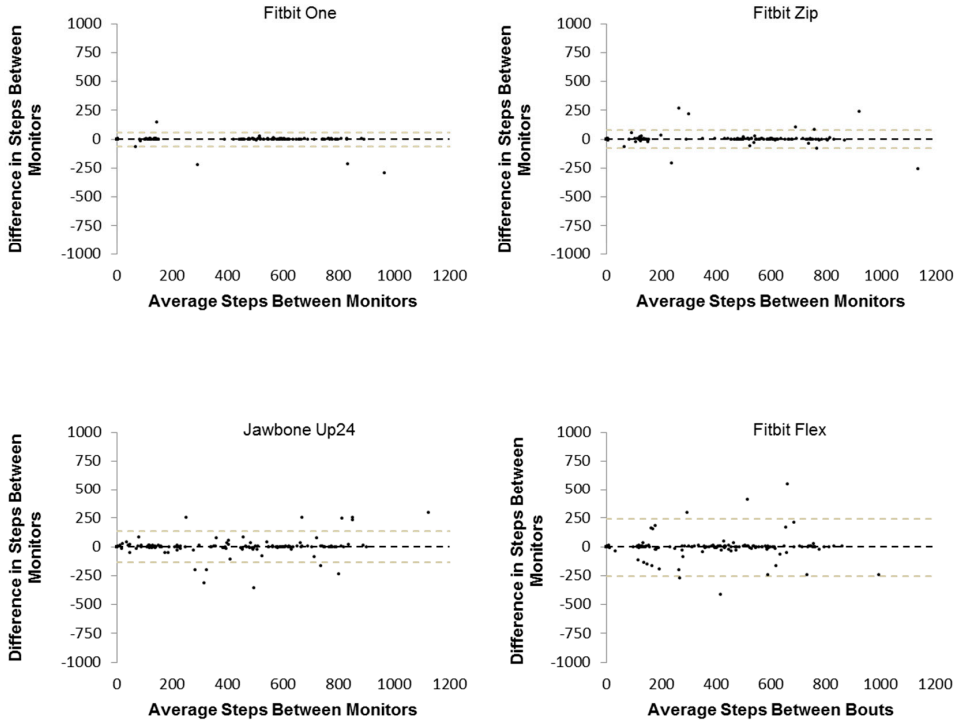

Figure 4. Bland-Altman plots with 95% limits of agreement created using the inter-monitor step data from all activities completed by each participant.

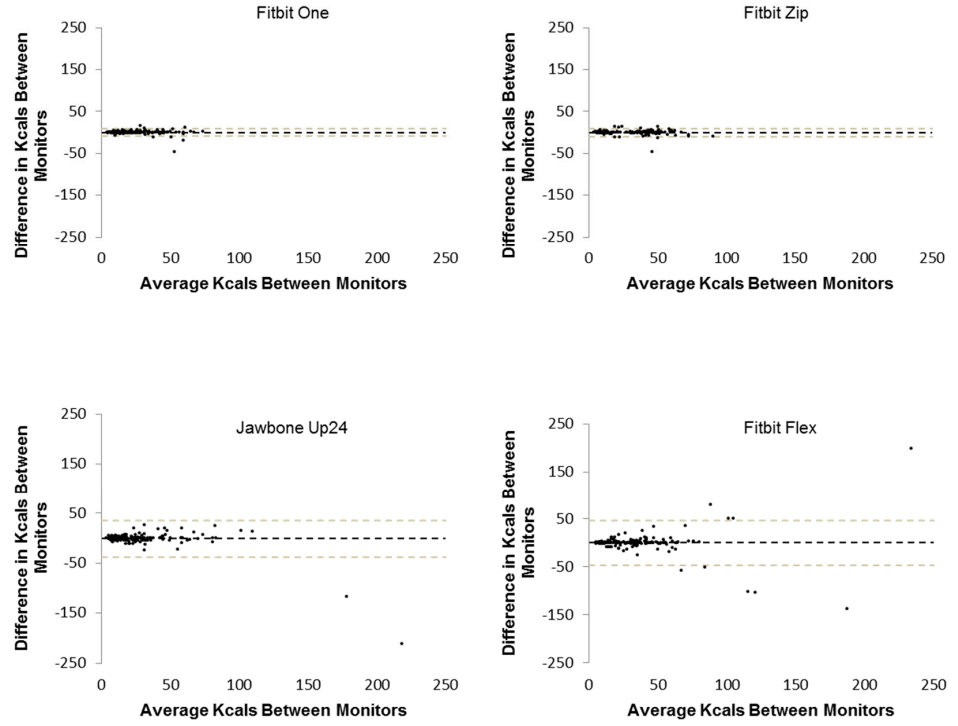

Figure 5. Bland-Altman plots with 95% limits of agreement created using the inter-monitor kilocalorie (kcals) data from all activities completed by each participant.

6.3. Inter-Monitor Reliability: Free-Living Setting

Each pair of CPAMs (wrist- or hip-worn) was worn by fifteen participants. A small percentage of step (3.6%), kcal (0.0%), and active minute (5.0%) were removed per exclusion criteria. Minimum wear time was not mandated; however, mean wear time was 5.7 ± 3.8 h. Correlations for steps, kcals, and active minutes for all CPAMs are shown in Figure 6. Most CPAMs had high inter-monitor reliability for all variables, except for kcals for the FO (low), active minutes for the FZ (moderate), and steps/kcals for the FF (moderately high). The abnormally low FO kcal and FZ active minutes correlations are attributable to infrequent outliers illustrated in Figures 7–9. MAD and MPD data paralleled data collected in the laboratory setting; that is, wrist-worn CPAMs displayed greater (worse) MAD and MPD data compared to the hip-worn CPAMs. JU had the highest step and active minute MAD and MPDs while the FF had the largest MAD and MPD for kcals.

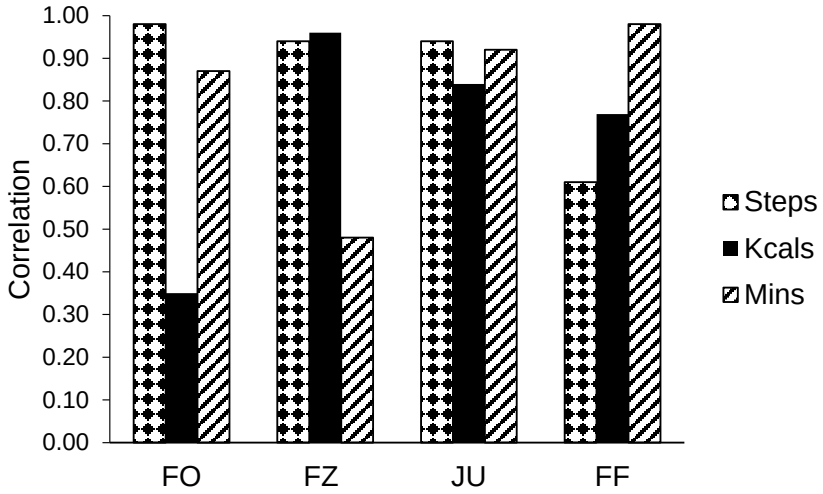

Figure 6. Pearson correlations (inter-monitor reliability) of the free-living data. Kcals = Calories. Mins = active minutes.

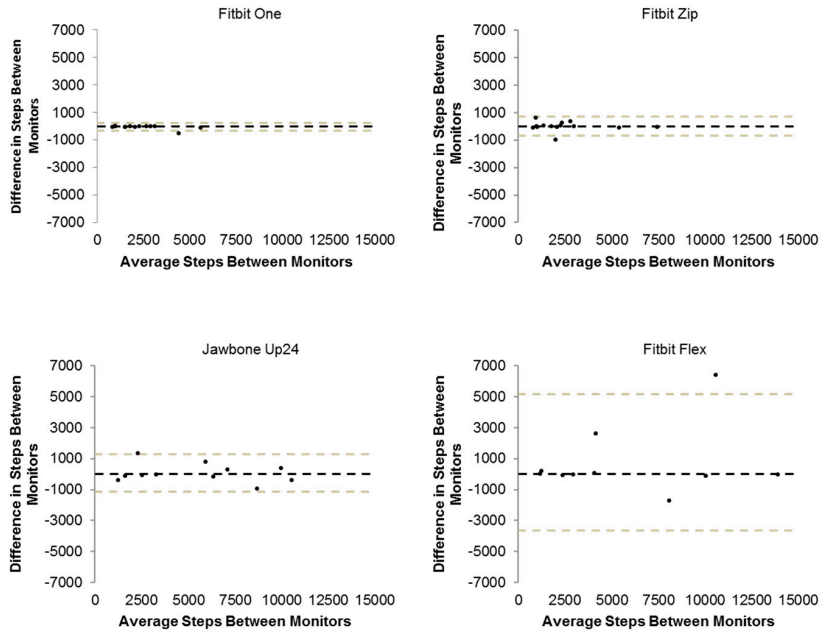

Figure 7. Bland-Altman plots with 95% limits of agreement created using the free-living steps data from each participant.

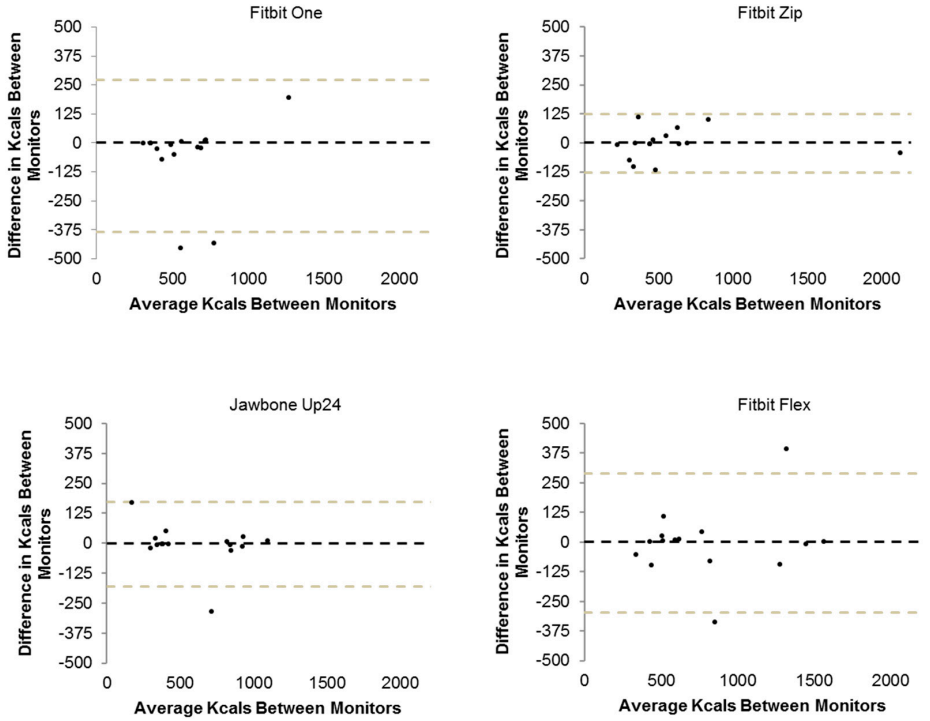

Figure 8. Bland-Altman plots with 95% limits of agreement created using the free-living kilocalories (kcals) data from each participant.

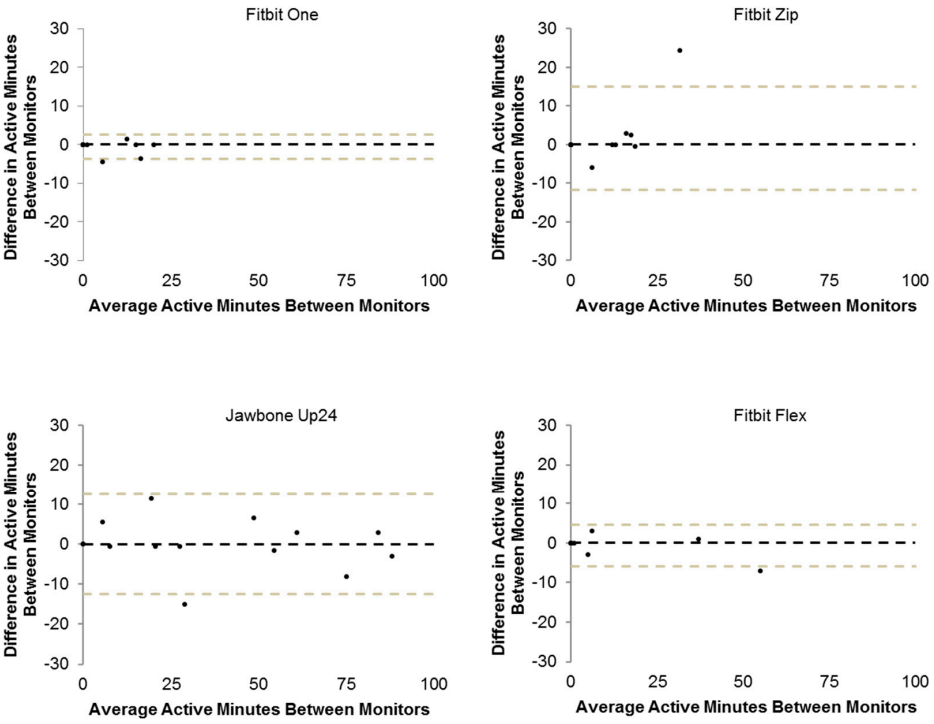

Figure 9. Bland-Altman plots with 95% limits of agreement created using the free-living active minutes data from each participant.

7. Discussion

This study found that all CPAMs had high intra-monitor reliability (≥0.80) for steps in a laboratory setting; however, the hip-worn CPAMs were significantly more reliable than the wrist-worn CPAMs. The ICCs in the present study are higher than those found by Kooiman et al. [8] who determined intra-monitor reliability for steps using the FF, JU, and FZ; their ICCs were 0.81, 0.83, and 0.90, respectively compared to 0.89, 0.88, and 0.99 in our study also using a laboratory setting [8]. Discrepancies between the studies could be attributable to differences in activity protocols. Kooiman et al. [8] used a single treadmill walking activity, whereas the present study used seven different activities, including both ambulatory (e.g., walking) and non-ambulatory (e.g., typing) tasks. The larger number and greater variety of activities used in our study builds upon preexisting CPAM literature and advances our understanding on how these devices perform during free-living activity. While no other studies have investigated the intra-monitor reliability of the FO for steps or any CPAMs to estimate kcals, we found lower reliability for kcal estimated than for step estimates, both in terms of lower correlations but also higher MPD. Our findings show consistently high intra-monitor reliability, especially for step estimates, with a variety of CPAMs and activities.

High correlations (≥0.80) were also observed in the inter-monitor reliability analyses for all CPAMs when estimating steps but only for the hip-worn CPAMs when estimating kcals; wrist-worn CPAMs had moderately-high correlations for kcal estimations. When examining CPAM validity, most studies show higher accuracy for step estimations than kcal estimates [7,16,17]. Therefore, available evidence suggests that step estimations from CPAMs are both more reliable and more valid than kcal estimations. The correlations obtained in the present study were comparable to those reported by Diaz et al. for steps (0.97 vs. 0.99) and kcals (0.94 vs. 0.97), as well as those of Takacs et al. [11] for steps (0.96 vs. 1.00), respectively [8,10]. It is important to note the consistently high correlations across various activity protocols indicating that reliability remains high even with the inclusion of a variety of activities, contrasting validity research where inclusion of diverse activities lowers CPAM validity [18].

CPAM correlation point estimates from the free-living portion of this study were comparable to, or lower than, those found during the laboratory portion. Most correlations were moderately high, although there were four instances when CPAM performance failed to meet the moderately high threshold. These instances included the FO for kcals, the FZ for active minutes, and the FF for steps and active minutes. The JU was the only CPAM whose correlations met the moderately high or greater criteria for all PA variables. A case study examining inter-monitor reliability of 10 Fitbit Ultra devices in an eight-day free-living trial found considerably higher reliability coefficients (0.995–1.000) for daily step counts than our study [12]. The Fitbit Ultra is a hip-worn CPAM, which partially explains the strong reliability found in their study. Additionally, only a portion of the day was spent in a free-living setting in our study, there was likely smaller variability in the data collected subsequently causing lower correlation coefficients than seen in the work of Dontje et al [12]. However, recent works have illustrated that CPAM's underestimate PA in free-living settings [14] and that the variability of these estimations is not consistent between CPAM models [13]. Collectively, available research suggests high or moderately-high reliability for most CPAMs and most dependent variables tested in free-living settings, supporting the use of these CPAMs during field-based PA monitoring [12].

While all CPAMs in the present study yielded moderately-high to high intra- and inter-monitor reliability in the laboratory, the hip-worn CPAMs (FO and FZ) had higher reliability than the wrist-worn CPAMs (JU and FF), both in terms of correlations as well as smaller (better) intra- and inter-monitor MAD and MPD values and generally narrower (better) 95% limits of agreement when examined using Bland-Altman plots. Given the greater variability and higher accelerations of arm movement compared to hip movement during basic tasks, these results were expected. However, wrist-worn activity monitors have better user compliance than hip-worn monitors [2,19,20]. Additionally, there are a greater number of wrist-worn CPAMs than hip-worn CPAMs on the market suggesting that wrist-worn CPAMs may be the more popular models. Accordingly, the choice of CPAM placement (wrist vs. hip) may depend on the importance of optimal reliability vs. optimal compliance and comfort.

All CPAMs in the present study collect and interpret PA data based upon accelerometer-based sensors within the device. More recently, manufacturers have produced CPAMs which incorporate variables, such as heart rate or other variables (e.g., skin temperature), into their algorithms (e.g., Apple Watch and Fitbit Charge). Indeed, a recent study showed these multi-sensor CPAMs showed improved energy expenditure estimations compared to single-sensor CPAMs [13]. As multi-sensor CPAMs become more common, the reliability of their newer variables (e.g., heart rate) and the influence of newer variables on other preexisting variables (e.g., kcals) should be investigated as there is likely crosstalk between sensors.

This study did not evaluate CPAM validity, but the relationship between CPAM validity and reliability is worth considering. A recent review article by Evenson et al. [7] reported results from over 20 validity and reliability studies, finding high validity and inter-monitor reliability for steps using treadmill-oriented protocols. Evenson also noted lower monitor validity during non-ambulatory activities and when the CPAMs were used in a free-living setting [7]. In contrast, our study found high or moderately high intra- and inter-monitor reliability across a variety of ambulatory and non-ambulatory activities in both laboratory and in free-living settings. Thus, available evidence suggests that CPAM reliability may be stronger than CPAM validity; in other words, CPAMs may be more useful for tracking PA changes within an individual over time or comparing PA trends between individuals than assessing adherence to PA recommendations. This should be taken into consideration when determining the utility of CPAMs as tracking or intervention tools.

Weaker correlations were observed in isolated cases during the free-living data collection, likely as a result of large differences in predicted activity in a few participants (Figures 7–9). Interestingly, there is a noticeable discrepancy between these correlations and their respective MAD and MPD values. While these results seemingly contradict one another, the large differences which significantly impacted the correlations are not as influential in an analysis of the median absolute and percent differences because median values are not sensitive to outliers. The robustness of median values (compared to means) allows for better interpretation of differences between monitors and is supported by its use in previous research [15]. Nevertheless, the large differences and data removed per the 'cleaning' process are worth noting. It is the authors' impression that the artificial laboratory analysis and subsequently frequent uploading of CPAM data may have introduced some of these data (e.g., Figure 5). CPAMs are likely not intended to be updated in five-minute intervals over an extended period of time; thus, these errors could be attributable to application lag. It is worth noting, though, that these devices may have occasional errors while updating one or more variables. This may have contributed to some of the instances where the variables were actually lower at the end of the day than the beginning of the day (which is not physiologically possible). Additionally, some of the sporadic large differences seen in the free-living data (e.g., a difference of >6000 steps for the Fitbit Zip; differences of >375 kcals for the Fitbit One and Flex; differences of >10 active minutes for the Fitbit Zip and Jawbone Up24) may be attributable to occasional data loss during updating. Issues with updating the devices and/or associated applications are worth noting as they quantitatively lower reliability and may necessitate data screening or removal rules to be introduced.

A limitation of this study design was the relatively short duration (five minutes) of the laboratory activities which did not permit the analysis of active minutes (require at least 10-min bouts for Fitbit monitors). The abbreviated activity times may have also contributed to relatively frequent failure of CPAMs and/or their related applications to update properly resulting in bad data (removed from analysis). Additionally, sweeping was the only non-ambulatory, non-sedentary activity in the present study, which limits understanding of CPAM reliability during these types of activities. Mean wear time during the free-living portion of the study is also a limitation as it resulted in low data variability and limited options for statistical analyses for these data. To this, the limited wear time of these monitors in the free-living setting did not permit a statistical comparison between laboratory-based and free-living performance. Furthermore, a washout period was not utilized between laboratory and free-living segments of the study. This limitation introduces a source of variability such that participants may

have modified their free-living behavior having completed the laboratory protocol earlier in the day. However, the laboratory activity protocol included a variety of activities not previously assessed in reliability studies (e.g., sweeping and reading) strengthens the present study. By including these activities of daily living, our results better reflect the performance of these monitors to across a variety of activities likely to be performed during a typical day. Second, this study included both laboratory and free-living aspects, which provides a more developed assessment of CPAM performance compared to studies without a free-living component.

In conclusion, these CPAMs provide reliable estimations of most PA variables in the laboratory; however, their reliability declines in a free-living setting. This may be attributable to small discrepancies between estimations being amplified as a result of increased wear time. Nonetheless, these findings suggest that certain CPAMs can provide reliable estimations of PA, especially steps taken, in a laboratory setting and possibly in free-living.

Acknowledgments: The authors would like to thank Reem Hindi and Gabriela Torres for assistance with data collection and subject recruitment.

Author Contributions: Joshua M. Bock, Leonard A. Kaminsky, Matthew P. Harber, and Alexander H. K. Montoye conceived and designed the study. Joshua M. Bock and Alexander H. K. Montoye were responsible for collecting data. Joshua M. Bock was responsible for data analysis. Joshua M. Bock and Alexander H. K. Montoye wrote the initial manuscript. Joshua M. Bock, Leonard A. Kaminsky, Matthew P. Harber, and Alexander H. K. Montoye revised and prepared the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Troiano, R.P.; Berrigan, D.; Dodd, K.W.; Masse, L.C.; Tilert, T.; McDowell, M. Physical activity in the United States measured by accelerometer. *Med. Sci. Sports Exerc.* **2008**, *40*, 181–188. [CrossRef] [PubMed]
2. Troiano, R.P.; McClain, J.J.; Brychta, R.J.; Chen, K.Y. Evolution of accelerometer methods for physical activity research. *Br. J. Sports Med.* **2014**, *48*, 1019–1023. [CrossRef] [PubMed]
3. Center for Disease Control Physical Activity Data and Statistics. Available online: http://www.cdc.gov/physicalactivity/data/ (accessed on 11 April 2017).
4. Wareham, N.J.; Rennie, K.L. The assessment of physical activity in individuals and populations: Why try to be more precise about how physical activity is assessed? *Inter. J. Obes.* **1998**, *22*, S30–S38.
5. Inside wearables Part 2. Available online: http://digitalintelligencetoday.com/wp-content/uploads/2015/11/2014-Inside-Wearables-Part-2-July-2014.pdf (accessed on 11 April 2017).
6. Worldwide Wearables Market Increases 67.2% Amid Seasonal Retrenchment, According to IDC. Available online: http://www.idc.com/getdoc.jsp?containerId=prUS41284516 (accessed on 11 April 2017).
7. Evenson, K.R.; Goto, M.M.; Furberg, R.D. Systematic review of the validity and reliability of consumer-wearable activity trackers. *Int. J. Behav. Nutr. Phys. Act.* **2015**, *12*, 159–181. [CrossRef] [PubMed]
8. Kooiman, T.J.M.; Dontje, M.L.; Sprenger, S.R.; Krijnen, W.P.; van der Schans, C.P.; de Groot, M. Reliability and validity of ten consumer activity monitors. *BMC Sport Sci. Med. Rehabil.* **2015**, *7*, 24–35. [CrossRef] [PubMed]
9. Diaz, K.M.; Krupka, D.J.; Chang, M.J.; Peacock, J.; Ma, Y.; Goldsmith, J.; Schwartz, J.E.; Davidson, K.W. Fitbit: An accurate and reliable device for wireless physical activity tracking. *Int. J. Cardiol.* **2015**, *185*, 138–140. [CrossRef] [PubMed]
10. Mammen, G.; Gardiner, S.; Senthinathan, A.; McClemont, L.; Stone, M.; Faulkner, G. Is this bit fit? Measuring the quality of the Fitbit step-counter. *Health Fit. J. Can.* **2012**, *5*, 30–39.
11. Takacs, J.; Pollock, C.L.; Guenther, J.R.; Bahar, M.; Napier, C.; Hunt, M.A. Validation of the Fitbit One activity monitor device during treadmill walking. *J. Sci. Med. Sport* **2014**, *17*, 496–500. [CrossRef] [PubMed]
12. Dontje, M.L.; de Groot, M.; Lengton, R.R.; van der Schans, C.P.; Krijnen, W.P. Measuring steps with the Fitbit activity tracker: An inter-device reliability study. *J. Med. Eng. Tech.* **2015**, *39*, 286–290. [CrossRef] [PubMed]
13. Chowdhury, E.A.; Western, M.J.; Nightingale, T.E.; Peacock, O.J.; Thompson, D. Assessment of laboratory and daily energy expenditure estimates from consumer multi-sensor physical activity monitors. *PLoS ONE* **2017**, *24*, e0171720. [CrossRef] [PubMed]

14. Murakami, H.; Kawakami, R.; Nakae, S.; Nakata, Y.; Ishikawa-Takata, K.; Tanaka, S.; Miyachi, M. Accuracy of wearable devices for estimating total energy expenditure: Comparison with metabolic chamber and doubly labeled water method. *JAMA Intern. Med.* **2016**, *176*, 702–703. [CrossRef] [PubMed]
15. Ferguson, T.; Rowlands, A.V.; Olds, T.; Maher, C. The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: A cross-sectional study. *Int. J. Behav. Nutr. Phys. Act.* **2015**, *12*, 42–51. [CrossRef] [PubMed]
16. Safrit, M.J.; Wood, T.M. *Introduction to Measurement in Physical Education and Exercise Science*; Mosby: St. Louis, MO, USA, 1995.
17. Swartz, A.M.; Strath, S.J.; Bassett, J.R.; O'Brien, W.L.; King, K.A.; Ainsworth, B.E. Estimation of energy expenditure using CSA accelerometers at the hip and wrist sites. *Med. Sci. Sports Exerc.* **2000**, *32*, S450–S456. [CrossRef] [PubMed]
18. Nelson, M.B.; Kaminsky, L.A.; Dickin, D.C.; Montoye, A.H.K. Validity of consumer-based physical activity monitors. *Med. Sci. Sport Exer.* **2016**, *48*, 1619–1628. [CrossRef] [PubMed]
19. Kamada, M.; Shiroma, E.J.; Harris, T.B.; Lee, I.M. Comparison of physical activity assessed housing hip- and wrist-worn accelerometers. *Gait Posture* **2016**, *44*, 23–28. [CrossRef] [PubMed]
20. Fairclough, S.J.; Noonan, R.; Rowlands, A.V.; van Hees, V.; Knowles, Z.; Boddy, L.M. Wear compliance and activity in children wearing wrist- and hip-mounted accelerometers. *Med. Sci. Sports Exerc.* **2016**, *48*, 245–253. [CrossRef] [PubMed]