Measurement in Physical Education and Exercise Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hmpe20

Agreement between Fitbit and ActiGraph Estimates of Physical Activity in Young Children

Michael D Schmidt, Stephen L. Rathbun, Zhixuan Chu, Benjamin D. Boudreaux, Lindsay Hahn, Eric Novotny, Kyle Johnsen & Sun Joo (Grace) Ahn

MEASUREMENT IN PHYSICAL EDUCATION AND EXERCISE SCIENCE
https://doi.org/10.1080/1091367X.2022.2125319

Agreement between Fitbit and ActiGraph Estimates of Physical Activity in Young Children

Michael D Schmidt [a], Stephen L. Rathbun [b], Zhixuan Chu [b], Benjamin D. Boudreaux [a], Lindsay Hahn [c], Eric Novotny [d], Kyle Johnsen [e], and Sun Joo (Grace) Ahn [d]

[a]Department of Kinesiology, University of Georgia, Athens, Georgia, USA; [b]Department of Epidemiology and Biostatistics, University of Georgia, Athens, Georgia, USA; [c]Department of Communication, University at Buffalo, State University of New York, Buffalo, New York, USA; [d]Grady College of Journalism & Mass Communication, University of Georgia, Athens, Georgia, USA; [e]College of Engineering, University of Georgia, Athens, Georgia, USA

ABSTRACT

Physical activity (PA) estimates from the Fitbit Flex 2 were compared to those from the ActiGraph GT9X Link in 123 elementary school children. Steps and intensity-specific estimates of PA and 3-month PA change were calculated using two different ActiGraph cut-points (Evenson and Romanzini). Fitbit estimates were 35% higher for steps compared to the ActiGraph. Fitbit and ActiGraph intensity-specific estimates were closest for sedentary and light PA, while estimates of moderate and vigorous PA varied substantially depending upon the ActiGraph cut-points used. Spearman correlations between device estimates were higher for steps (r_s = .70) than for moderate (r_s =.54 to .55) or vigorous (r_s =.29 to .48) PA. There was low concordance between devices in assessing PA changes over time. Agreement between Fitbit Flex 2 and ActiGraph estimates may depend upon the cut-points used to classify PA intensity. However, there is fair to good agreement between devices in ranking children's steps and MVPA.

KEYWORDS
Validity; fitness trackers; steps; change measurement; youth

Introduction

Fitbit devices were introduced in 2008 as a commercially available tool for individuals to track their daily steps, calories, and sleep (Evenson et al., 2015). The first models (e.g., Ultra, One, Zip) were clip-on devices that could be worn at the waist, pocket, or bra, but all models introduced after 2013 are wrist-worn devices. Early wrist-worn Fitbits, such as the Flex and the Flex 2, were mainly limited to estimating physical activity metrics. Currently available models (e.g., Ace 3, Charge 5, Inspire 2, Sense) include measurements of heart rate, blood oxygen levels, and sleep stages. Although originally intended for personal monitoring, Fitbit devices are increasingly being used to measure physical activity in research studies conducted in adults (Feehan et al., 2018) and children (Byun et al., 2018; Cradock et al., 2019; Evans et al., 2017). One reason for this increased use may be high compliance among research participants due to the familiarity and acceptability of Fitbit devices. For example, over 90% of women in the Nurse's Health Study 3 (NHS) Mobile Health Substudy wore their assigned Fitbit device for 10+ hr/d on 5 or more days during the first week of the study (Fore et al., 2020). Other factors facilitating the increased research use of Fitbits may include the ability to obtain minute-by-minute activity metrics, and lower researcher costs and burden relative to research-grade activity monitors (Feehan et al., 2018).

To date, Fitbit devices have been used to estimate and compare physical activity levels in population subgroups and to estimate changes in physical activity over time. Because studies may use Fitbit devices as the primary physical activity measurement tool, there is a need to accumulate evidence supporting the validity of Fitbit estimates for these uses in real-world settings and among diverse populations. Toward this goal, Evenson et al. (2015) reviewed 16 studies that evaluated the validity of one or more physical activity metrics (e.g., steps/day, moderate to vigorous physical activity [MVPA] min/day) from several Fitbit models in adult samples. Most of the studies compared Fitbit step estimates during laboratory-based activities to observed or accelerometer measured steps and found strong correlations (≥ .80), with Fitbits being slightly more accurate when worn on the hip compared to the wrist. A subsequent review by Feehan et al. (2018) updated the evidence and included 13 studies that estimated the validity of Fitbit step estimates in free-living conditions.

CONTACT Michael D Schmidt schmidtm@uga.edu Department of Kinesiology, University of Georgia, 115J Ramsey Center, 330 River Rd, Athens, Georgia 30602, USA

In healthy young adults, Fitbit step estimates were within ±10% of research-grade accelerometer estimates in nine of these studies, although the Fitbit tended to overestimate steps compared to the criterion measures by approximately 8%. Averaged across seven studies, Fitbit estimates of free-living MVPA were 85% higher than ActiGraph estimates.

In comparison to the robust literature for adults, few studies have examined the validity of Fitbit devices in school-age youth, with only five studies identified that examined the validity of step and/or MVPA estimates among children. Three studies (Godino et al., 2020; Hao et al., 2021; Kang et al., 2019) evaluated wrist-worn models (Charge HR and Flex) and each of these studies compared the accuracy of the Fitbit to research-grade measures during standardized laboratory- and field-based activities. The accuracy of Fitbit step estimates varied by model and protocol, with the Flex overestimating steps by 21% compared to a wrist-worn ActiGraph GT3X+ (mean absolute percent error [MAPE] = 21.9%; Hao et al., 2021), while the Charge HR underestimated ambulation by 11.8 steps/min compared to video observation (MAPE = 9.9%; Godino et al., 2020). The Charge HR also showed a moderate-to-high level of accuracy (71–85%) in classifying minutes spent in MVPA (Godino et al., 2020; Kang et al., 2019).

While three studies have examined the validity of wrist-worn devices in school-aged youth, none were conducted during free-living conditions. This is an important limitation as children are likely to spend most of their time in non-structured physical activities, which may be more difficult to accurately capture in the laboratory. Wrist-worn devices may be especially prone to error during these types of activity if they consist of arm actions without corresponding whole-body movements. These studies were also conducted in relatively small samples (19 to 59 children) and included only children aged 8 years and above. Further, prior studies in youth have not compared the ability of Fitbit devices to capture changes in activity over time relative to research-grade devices. This is an important omission given that Fitbit devices are being used to track changes in steps and MVPA in research studies (Buchele Harris & Chen, 2018; Evans et al., 2017; Hayes & Van Camp, 2015) and cross-sectional findings may not reflect the ability of Fitbit devices to accurately estimate longitudinal changes in PA.

To address the noted research gaps, this study compared estimates of free-living physical activity from a wrist-worn Fitbit Flex 2 to estimates from a waist-worn ActiGraph GT9X Link in a diverse sample of elementary school children. Two alternative sets of cut-points were used to derive intensity-specific PA estimates from the ActiGraph in order to examine the effect of cut-point selection on the level of agreement between Fitbit and ActiGraph devices. A secondary aim was to compare Fitbit and ActiGraph estimates of physical activity change over a 3-month period. We chose to use a waist-worn ActiGraph as the criterion measure for this study as this wear location has demonstrated superior accuracy in classifying physical activity intensity compared to a wrist-worn ActiGraph (Migueles et al., 2017) and is the most common location of ActiGraph wear in studies evaluating consumer wearable devices (Gorzelitz et al., 2020).

Methods

Study design and participants

Data for this analysis were collected in 2018–2019 during the first year of an ongoing technology-based intervention intended to increase physical activity participation among 6–11-year-old children attending 12 YMCA after-school programs in the metro Atlanta region (Georgia, USA). Details of the parent study intervention, including participant recruitment strategies and compensation, have been published elsewhere (Hahn et al., 2020). Parent consent and child assent were obtained for all study participants, and study procedures were approved by the university's institutional review board.

Briefly, children in both the intervention and control arms of the study were asked to wear a Fitbit Flex 2 activity monitor continuously for one academic year, and a subsample of children was invited to wear an ActiGraph GT9X activity monitor for seven continuous days at four points throughout the year (baseline, and 3, 6, and 9 months after the intervention's start). Researchers also measured children's body composition and administered self-report surveys to children at these four time points. The current study reports data obtained from a subsample of children who wore both a Fitbit Flex 2 and an ActiGraph GT9X at the baseline and 3-month measurement periods. These periods were selected to examine the agreement between devices in estimating physical activity changes during the time period where the largest changes in children's activity levels were expected to occur due to the intervention.

Measures

Fitbit flex 2

To obtain daily physical activity information, children were instructed to continuously wear a Fitbit Flex 2 (Fitbit Inc., San Francisco, CA) activity tracker on their non-dominant wrist (default setting) except during device charging and when wear was prohibited during organized sports. This device uses data from an internal 3-axis accelerometer to estimate steps and minutes accumulated at three physical activity intensity levels (lightly active, fairly active, and very active intensity). Although the algorithms used to determine physical activity intensity are proprietary, it has been assumed that these categories approximately correspond to light (1.6–2.9 METs), moderate (3.0–5.9 METs), and vigorous (≥6.0 METs) physical activity. All other minutes were assumed to be of sedentary intensity (≤1.5 METs). Periodically, minute-by-minute device data were automatically uploaded to the Fitbit online platform and subsequently downloaded through an application-programming interface. Upon enrollment, parents were asked to give researchers permission to view and record their child's Fitbit data for the duration of the study.

In the absence of a standard procedure, daily Fitbit wear times were estimated using a modification of the Choi algorithm (Choi et al., 2011) developed for use with ActiGraph accelerometer data. Specifically, 90 minutes of consecutive zero step counts, instead of accelerometer counts, were used to identify periods of probable non-wear. Similar methods have been used to assess Fitbit wear times in studies of both children (Byun et al., 2018; Cradock et al., 2019) and adults (Collins et al., 2019; Dominick et al., 2016).

ActiGraph GT9X

In addition to the Fitbit, at each of the four measurement periods up to 20 children at each after-school program were fitted with ActiGraph GT9X accelerometers (ActiGraph Corp., Pensacola, FL) attached to an elastic belt and positioned at the mid-axillary line of the right hip. The devices were set to display only the current time. Children were verbally instructed to wear the device during all waking hours, except for water-based activities, for seven consecutive days. Parents received detailed written instructions and text message reminders during the wear-period and were asked to record the times the device was worn daily and reasons for non-wear.

For this analysis, ActiGraph devices were equipped with Firmware version 1.7.1 and Actilife software version 6.13.4 was used to initialize and download data in 1-minute epochs to match the data output from the Fitbit devices. Accelerometer count values during each epoch were used to classify each minute into a physical activity intensity category (i.e., sedentary, light, moderate, vigorous) using two alternative sets of age-appropriate cut-points (Evenson et al. (2008); (Romanzini et al., 2014)). Briefly, the Evenson cut-points were developed in 5–8 year-old children and use only vertical axis counts to determine PA intensity (Sedentary: 0–100 counts per minute (cpm); Light: 101–2295 cpm; Moderate: 2296–4011 cpm; Vigorous: ≥4012 cpm). The Romanzini cut-points were developed in 10–15 year-old youth and are based on the vector magnitude of count values from triaxial ActiGraph models (Sedentary: 0–720 counts per minute (cpm); Light: 721–3027 cpm; Moderate: 3028–4447 cpm; Vigorous: ≥4448 cpm).

To match the approach used with the Fitbit data, periods of device wear were estimated using a modified version of the Choi algorithm (Choi et al., 2011), where periods with consecutive zero step values for 90 minutes or longer were classified as non-wear. To examine the impact of applying the Choi algorithm to steps, we compared ActiGraph wear time estimates generated from steps versus estimates generated from counts using data from 183 children collected during the first four complete days of baseline data collection. Wear times based on ActiGraph steps were only 2.4% lower than mean estimates derived from ActiGraph counts, and estimated wear times did not differ between the approaches on 60% of days. The correlation between the wear time estimates was .989.

ActiGraph reliability and validity have been consistently demonstrated (Trost et al., 2005) and it has been shown to correlate reasonably with activity energy expenditure measured by doubly labeled water (Plasqui et al., 2013; Plasqui & Westerterp, 2007).

Body size

Children's height and weight were obtained at the baseline, 3-, 6-, and 9-month measurement periods. All measures were taken in light clothing and with shoes and socks removed. Height was measured to the nearest .1 cm using a Hopkins Road Rod portable stadiometer (Caledonia, MI) and body weight (to the nearest .1 kg) was measured using a Tanita BF-689 Children's Body Fat Monitor (Arlington Heights, IL). Age- and sex-specific BMI percentiles were used to classify children as overweight (85th to <95th percentile) and obese (≥95th percentile; Barlow & Expert, 2007).

Demographics

In the parent's self-report surveys at baseline, information was obtained about the parents' socio-economic status and about their child's age, gender, and race/ethnicity.

Data processing

For this analysis, the available Fitbit and ActiGraph data were first restricted to those minutes when both devices were worn and then to days with at least 8 hr of concordant wear per day. As the goal of this study was not to estimate the habitual physical activity levels of the study participants, but to compare the estimates from two monitors during concordant wear periods, we included all children with at least one 8 hr day of concordant wear. Daily estimates of physical activity were then divided by the duration (in hours) the devices were worn to derive physical activity metrics per hour of device wear. Daily Fitbit and ActiGraph estimates of activity were then averaged for each child. To compare Fitbit and ActiGraph estimates of physical activity change, separate estimates were calculated using data collected at baseline and at the 3-month data collection period. Changes in Fitbit and ActiGraph physical activity metrics were derived by subtracting average baseline values from average 3-month values for each device.

Statistical analysis

Descriptive statistics and histograms were obtained for all continuous variables at both the day and subject level to assess normality and to screen for data anomalies. Scatterplots were used to examine the nature of the relationship between corresponding Fitbit and ActiGraph estimates of activity and to identify potentially influential data points. Because of the presence of influential outliers and the non-normality of some variables, Spearman correlations were used to examine the strength of the monotonic relationships between corresponding physical activity estimates, and 95% confidence intervals (CI) were calculated based on Fisher transformation.

Equivalence tests were performed to assess the agreement between the Fitbit and ActiGraph estimates of physical activity. Because our data were paired, we calculated the mean difference between the paired observations for each activity estimate along with the 90% confidence interval (CI) of the difference. This interval was compared to an equivalence region of ±10% of the mean ActiGraph estimates (generated separately using the Evenson and the Romanzini cut-points). The Fitbit estimate is considered significantly equivalent to an ActiGraph estimate with $p < .05$ when the 90% CI of the difference is completely within the designated equivalence region (Dixon et al., 2018). Mean absolute percent deviations (MAPD) with 95% CIs were calculated as an additional measure of agreement between the Fitbit estimates and the two sets of ActiGraph estimates. MAPD values were also calculated to assess the agreement between ActiGraph PA estimates generated using the Evenson versus Romanzini cut-points. Because most studies require a minimum number of valid wear days for inclusion in study analyses, key comparisons were repeated after restricting our sample to children with at least four valid days of wear for both devices (n = 86).

Physical activity estimates were separately created for data collected at baseline and at 3-months into the data collection. Changes in Fitbit and ActiGraph estimates of physical activity (3-month minus baseline) were calculated in 40 children who simultaneously wore both devices for at least one valid day (i.e., 8+ hr) and these changes were compared using Spearman correlations. A sensitivity analysis was run in a subsample of 21 children who had 3+ days of concordant wear at both baseline and 3-months. All analyses were conducted using SAS version 9.4 (SAS Institute, Inc., Cary, NC) and statistical significance was set at an alpha level of .05.

Results

During the overlapping wear periods, there were 660 days where both the ActiGraph and Fitbit device were simultaneously worn for at least 8 hr. A total of 123 children had at least one valid measurement day with both devices [mean (SD) wear time of 12.2 (1.4) hr/day], with 86 children having ≥4 valid days [mean (SD) wear time of 12.3 (1.2) hr/day]. As detailed in Table 1, children with at least one valid day were diverse in sex, race/ethnicity, and parental income [55.3% male, 61.7% non-white and non-Hispanic, and 31.7% earning <\$50,000 (among those reporting)]. A majority of children were either overweight (41.0%) or obese (23.8%) based on BMI percentile cut-points.

Table 1. Descriptive characteristics of study participants (n = 123).

Demographics	M ± SD (range) or N (%)
Age	8.1 ± 1.4 (6–11)
Sex	
Male	68 (55.3%)
Female	55 (44.7%)
Race/Ethnicity	
White non-Hispanic	44 (35.8%)
Black non-Hispanic	44 (35.8%)
Hispanic	10 (8.1%)
South Asian	12 (9.8%)
Other [a]	5 (4.1%)
Not Reported	8 (6.5%)
Parent Income	
<\$50,000	32 (26.0%)
\$50,000-\$99,999	22 (17.9%)
≥\$100,000	47 (38.2%)
Not Reported	22 (17.9%)
Weight (kg)	33.1 ± 9.7 (17.5–61.7)
Height (cm)	133.1 ± 10.3 (106.9–161.0)
Body Mass Index (kg m^{-2})	18.3 ± 3.5 (13.1–28.9)
Body Mass Index (percentile)	68.7 ± 27.8 (.8–99.7)

Note. Weight, height, and body mass index values missing for one participant (n = 122).
[a]Includes n = 3 East Asian and n = 2 who reported "other" or "mixed" race.

Among children with at least one valid day, Fitbit step estimates were approximately 35% higher (228.4 steps/hr) than ActiGraph estimates and were deemed not equivalent as the 90% CI of these differences (199.8–257.1) extended well beyond the equivalence region (±65.9; Table 2). For an average subject with 14 hr of device wear, this would translate to about 3200 additional steps/day recorded by the Fitbit device. Equivalence testing also indicated non-equivalence between all Fitbit and ActiGraph intensity-specific measures based on the Evenson cut-points (AG_Evenson), as the 90% CIs of the mean differences extended outside of the designated equivalence regions. Using the Romanzini ActiGraph cut-points (AG_Romanzini), Fitbit estimates of sedentary time were found to be equivalent, with a mean difference of approximately .4 minutes per hour. However, all other AG_Romanzini intensity-specific estimates were not equivalent to the Fitbit estimates. Compared to the AG_Evenson estimates, Fitbit estimates were about 10% lower for moderate intensity and 37% higher for vigorous intensity activity (comparing medians) and overall MVPA estimates were 12% higher from the Fitbit. In contrast, Fitbit estimates were about 40% lower for moderate intensity and 65% lower for vigorous intensity activity (comparing medians) than AG_Romanzini estimates. Similar differences in step estimates from each device were observed when comparisons were restricted to children with at least 4 valid days. Differences between Fitbit and AG_Evenson estimates of MVPA were attenuated but remained non-equivalent.

Mean absolute percent deviations (MAPD) between Fitbit and ActiGraph estimates, as well as between AG_Evenson and AG_Romanzini estimates, are shown in Table 3. Compared to AG_Evenson estimates, Fitbit estimates of sedentary and light PA had the lowest MAPD values while vigorous PA estimates had the highest MAPD values. MAPD values were similar when Fitbit estimates were compared to AG_Romanzini estimates, although values were modestly lower for sedentary estimates (14.0 vs. 25.1) and substantially lower for vigorous PA estimates (66.0 vs. 264.0). Interestingly, when AG_Evenson and AG_Romanzini estimates were compared, MAPD values were similar in magnitude to those from the Fitbit and AG_Romanzini comparisons and lower than those from the Fitbit and AG_Evenson comparisons (except for light intensity). Similar patterns were observed when MAPD values were calculated only among children with at least 4 valid days.

Scatterplots generally revealed linear relationships between like estimates from the ActiGraph and Fitbit, although several potentially influential outliers (e.g., subjects with very high Fitbit steps but very low ActiGraph steps) were apparent and not due to any identifiable data processing errors. Therefore, Spearman correlations were calculated between like activity estimates from the ActiGraph and Fitbit (Table 4). The strongest correlations were observed between step estimates in both the full (r_s = .70) and restricted samples (r_s = .79). Fitbit intensity-specific estimates were more moderately correlated ActiGraph estimates and were similar in magnitude regardless of the ActiGraph cut-point used. The only exception was for vigorous intensity activity, in which

Table 2. Comparison of Fitbit and ActiGraph activity estimates, all subjects (n = 123) versus those with 4+ Valid days (n = 86).

			Evenson Cut-points		Romanzini Cut-points	
Activity Estimates	Fitbit (M ± SD)	Equivalence Region	Actigraph (M ± SD)	Mean Difference (90% CI)	Actigraph (M ± SD)	Mean Difference (90% CI)
All Subjects						
Steps (per hr)	886.5 ± 230.5	(-65.9, 65.9)	658.1 ± 175.8	228.4 (199.8, 257.1)	658.1 ± 175.8	228.4 (199.8, 257.1)
Intensity (min/hr)						
Sedentary	33.4 ± 3.9	(-2.85, 2.85)	28.5 ± 6.5	4.88 (4.00, 5.75)	33.0 ± 6.8	0.41 (-0.44, 1.27)
Light	23.1 ± 3.1	(-2.83, 2.83)	28.3 ± 5.8	-5.25 (-6.03, -4.48)	20.7 ± 5.1	2.37 (1.68, 3.06)
Moderate	2.2 ± 1.3	(-0.25, 0.25)	2.5 ± 1.1	-0.25 (-0.43, -0.08)	3.7 ± 1.4	-1.47 (-1.68, -1.27)
Vigorous [a]	1.3 ± 1.2	(-0.07, 0.07)	0.7 ± 0.5	0.63 (0.46, 0.80)	2.6 ± 1.4	-1.28 (-1.49, -1.07)
MVPA	3.5 ± 2.3	(-0.32, 0.32)	3.2 ± 1.4	0.38 (0.08, 0.68)	6.3 ± 2.6	-.2.76 (-3.12, -2.40)
Subjects with 4+ days						
Steps (per hr)	865.6 ± 202.4	(-65.9, 65.9)	658.8 ± 141.4	206.8 (184.5, 229.1)	658.8 ± 141.4	206.8 (184.5, 229.1)
Intensity (min/hr)						
Sedentary	33.4 ± 3.7	(-2.78, 2.78)	27.8 ± 5.1	5.57 (4.77, 6.38)	32.4 ± 5.6	0.99 (0.19, 1.79)
Light	23.3 ± 3.0	(-2.90, 2.90)	29.0 ± 4.4	-5.69 (-6.42, -4.95)	21.3 ± 4.1	2.00 (1.30, 2.70)
Moderate	2.1 ± 1.1	(-0.25, 0.25)	2.5 ± 1.0	-0.41 (-0.59, -0.23)	3.7 ± 1.1	-1.62 (-1.82, -1.43)
Vigorous [b]	1.2 ± 1.0	(-0.07, 0.07)	0.7 ± 0.4	0.52 (0.35, 0.69)	2.5 ± 1.3	-1.34 (-1.55, -1.13)
MVPA	3.3 ± 2.0	(-0.32, 0.32)	3.2 ± 1.3	0.11 (-0.19, 0.41)	6.2 ± 2.2	-2.96 (-3.32, -2.60)

Note. MVPA = moderate and vigorous intensity physical activity.
[a]Median (IQR) were .85 min/hr (.48, 1.91) for Fitbit, .62 min/hr (.33,.94) for ActiGraph (Evenson), and 2.49 min/hr (1.57, 3.44) for ActiGraph (Romanzini).
[b]Median (IQR) were .79 min/hr (.49, 1.90) for Fitbit, .63 min/hr (.34, .94) for ActiGraph (Evenson) and 2.48 min/hr (1.71, 3.34) for ActiGraph (Romanzini).

Table 3. Mean absolute percent deviations (MAPD) between Fitbit and ActiGraph activity estimates, all subjects (n = 123) versus those with 4+ Valid days (n = 86).

Activity Estimates	Fitbit vs. Evenson MAPD (95% CI)	Fitbit vs. Romanzini MAPD (95% CI)	Evenson vs. Romanzini MAPD (95% CI)
All Subjects			
Steps (per hr)	45.1 (30.6, 59.6)	N/A	N/A
Intensity (min/hr)			
Sedentary	25.1 (21.6, 28.5)	14.0 (12.1, 15.9)	13.7 (12.3, 15.1)
Light	24.1 (19.5, 28.6)	29.8 (16.5, 43.0)	40.1 (36.6, 43.6)
Moderate	43.0 (33.9, 52.1)	48.4 (42.2, 54.5)	33.9 (30.9, 36.9)
Vigorous	264.0 (120.5, 407.5)	66.0 (58.7, 73.4)	73.9 (71.8, 76.0)
MVPA	59.9 (44.8, 74.9)	50.6 (45.5, 55.7)	50.3 (48.6, 52.0)
Subjects with 4+ days			
Steps (per hr)	32.8 (28.6, 36.9)	N/A	N/A
Intensity (min/hr)			
Sedentary	24.6 (20.8, 28.5)	12.5 (10.3, 14.7)	14.0 (12.5, 15.6)
Light	20.6 (18.5, 22.7)	18.4 (14.2, 22.6)	38.0 (34.5, 41.4)
Moderate	33.9 (27.8, 39.9)	44.1 (39.1, 49.0)	33.8 (30.5, 37.0)
Vigorous	180.5 (107.4, 253.5)	61.0 (55.0, 67.0)	74.0 (72.2, 75.9)
MVPA	43.2 (33.2, 53.3)	49.1 (44.3, 53.9)	50.1 (48.4, 51.8)

Note. MVPA = moderate and vigorous intensity physical activity.

Table 4. Spearman correlations between Fitbit and ActiGraph estimates of similar activity metrics, all subjects versus those with 4+ Valid days.

Activity Estimates	All Subjects (n=123)	95% CI	Subjects with 4+ Days (n=86)	95% CI
Steps (per hr)	.70	.59, .78	.79	.69, .86
Fitbit vs. Evenson				
Time by Intensity (min/hr)				
Sedentary	.47	.32, .60	.46	.28, .61
Light	.49	.35, .62	.43	.24, .59
Moderate	.55	.42, .67	.50	.32, .64
Vigorous	.29	.12, .45	.37	.18, .54
MVPA	.52	.38, .64	.50	.32, .64
Fitbit vs. Romanzini				
Time by Intensity (min/hr)				
Sedentary	.54	.41, .66	.56	.39, .69
Light	.43	.28, .57	.41	.21, .57
Moderate	.54	.40, .65	.49	.31, .64
Vigorous	.48	.33, .60	.50	.33, .65
MVPA	.54	.40, .66	.51	.33, .65
Evenson vs. Romanzini				
Time by Intensity (min/hr)				
Sedentary	.86	.81, .90	.84	.76, .89
Light	.83	.77, .88	.80	.71, .87
Moderate	.77	.68, .83	.82	.74, .88
Vigorous	.83	.77, .88	.88	.81, .92
MVPA	.92	.88, .94	.94	.90, .96

Note. MVPA = moderate and vigorous intensity physical activity.

Fitbit estimates were more strongly correlated with AG_Romanzini estimates (r_s = .48 vs. r_s = .29). As expected, strong correlations were observed between AG_Evenson and AG_Romanzini intensity-specific estimates, with time spent in MVPA having the highest correlation (r_s = .92). When restricted to only those children with 4+ valid measurement days (n = 86), similar correlations were observed for all comparisons.

To examine the concordance of Fitbit and ActiGraph estimates of physical activity change, changes in steps and MVPA over a 3-month period were calculated for a subset of 40 children who simultaneously wore both devices for at least one valid day (i.e., 8+ hr) during the baseline and the 3-month follow-up periods. Mean (SD) 3-month changes in ActiGraph measured physical activity were small [−18.3 (193.7) steps/hr, +.18 (1.89) MVPA min/hr for AG_Evenson, and −.26 (2.62) MVPA min/hr for AG_Romanzini]. The Fitbit estimated change in steps was an average of 27.2 steps/hr higher than ActiGraph estimates (p = .53), while the Fitbit estimated change in MVPA was .23 min/hr lower (p = .64) than the AG_Evenson estimate and .20 min/hr higher (p = .73) than the AG_Romanzini estimate. Spearman correlations between Fitbit and ActiGraph change estimates were .34 (95% CI: .04, .59) for steps but only .05 (95% CI: −.26, .36) for AG_Evenson MVPA estimates and .14 (95% CI: −.18, .43) for AG_Romanzini MVPA estimates. Restricting the change analysis to 21 children with 3 + valid days for both devices at each time point did not substantively influence estimates of average change in steps or MVPA. However, the correlation between Fitbit and ActiGraph estimates of change was stronger for steps (r_s =.62; 95% CI: .26, .83), but not for MVPA (AG_Evenson r_s = .04, 95% CI: −.40, .46; AG_Romanzini r_s = .18, 95% CI: −.28, .57). For comparison, correlations between the AG_Evenson and AG_Romanzini MVPA change estimates were .88 (.79, .94) (n = 40) and .94 (.86, .98) (n = 21).

Discussion

The primary aim of this study was to compare free-living estimates of physical activity from a wrist-worn Fitbit Flex 2 to a waist-worn ActiGraph GT9X Link in a diverse sample of elementary school children. Two different ActiGraph youth cut-points were used to generate intensity-specific PA estimates, which allowed us to examine the effect of cut-point choice on the agreement between Fitbit and ActiGraph estimates. We observed average Fitbit estimates of steps to be approximately one-third higher than average ActiGraph estimates. The direction and magnitude of Fitbit versus ActiGraph differences in intensity-specific estimates were found to differ substantially depending upon the ActiGraph cut-points used. Equivalence testing confirmed that all Fitbit minus Actigraph differences fell outside the designated ±10% equivalence regions indicating non-equivalence, with the exception of Fitbit sedentary time, when compared to AG_Romanzini sedentary time. However, there was good rank-order agreement between step estimates from the two devices, and moderate agreement in estimates of MVPA regardless of the ActiGraph cut-points used.

To our knowledge, this is the first study to examine the validity of a wrist-worn Fitbit device in a diverse sample of school-aged youth during free-living conditions. The observed agreement between Flex 2 and ActiGraph GT9X estimates of MVPA is similar to that reported in a study of 27 preschool children who wore both the original Fitbit Flex and an ActiGraph GT3X+ over a 24-hr period (r = .58; MAPE = 55.7%; Byun et al., 2018). Other studies have evaluated the validity of wrist-worn Fitbit devices in children and adolescents during standardized laboratory- and field-based activities. For example, Godino et al. (2020) reported that the Charge HR underestimated steps by 11.8 steps/min compared to direct observation and overestimated energy expenditure by .55 METs/min compared to indirect calorimetry during a combination of 14 standardized activities completed by 59 boys and girls aged 9–11 years. In another study, Kang et al. (2019) compared the ability of the Charge HR and the ActiGraph GT3x+ (worn at the wrist) to correctly classify activity intensity in 43 children (aged 8–12 years) who completed 3-min bouts of 12 physical activities that ranged in intensity from sedentary to vigorous. Compared to estimates measured by a Cosmed portable metabolic unit, the Charge HR showed a moderate level of agreement (70.8%) for classifying MVPA, which was somewhat lower than the 81.8% agreement reported for the ActiGraph GT3x+ worn at the wrist.

One factor potentially contributing to the higher Fitbit estimates of steps and vigorous intensity PA observed in the current study (compared to AG_Evenson estimates) is that this device was worn on the non-dominant wrist, whereas the ActiGraph was worn on the waist. In a study of 188 children aged 9–12 years, ActiGraph GT3X+ estimates of MVPA were significantly higher from devices worn at the wrist, compared to devices simultaneously worn at the waist (McLellan et al., 2018). In addition, a recent study by Clevenger et al. (2019) found that Fitbit Flex 2 devices worn on the non-dominant wrist generated higher estimates of steps and activity minutes (5% and 20%, respectively) than Flex 2 devices simultaneously worn on the dominant wrist for 4 days in free-living conditions. Another potential factor contributing to the differences observed in the current study is the use of 1-minute epochs when classifying ActiGraph activity intensity levels. While this was done to match the epoch length of the Fitbit data, it is known that shorter epoch lengths (e.g., 15-s) are better able to capture sporadic bursts of higher intensity PA that are common in children (Migueles et al., 2017).

Most validation studies compare estimates from different measures at a single point in time. While sufficient to evaluate the ability of an instrument to correctly discriminate PA levels across individuals, such studies do not assess an instrument's ability to capture changes in PA behaviors over time, which is important when examining longitudinal trends or the effectiveness of a PA intervention. The few prior studies that compared estimates of PA change across measurement approaches have predominantly focused on the agreement between objective and subjective measures of change, finding poor concordance (Limb et al., 2019; Nicaise et al., 2014). We were unable to identify any prior comparisons of PA change estimates between commercial- and research-grade objective measures in youth or adults. In this study, we addressed this aim in a subsample of 40 children with sufficient data and observed no statistically significant differences in Fitbit and ActiGraph estimates of PA change over a 3-month period. However, only modest correlations were observed between estimates of step change from the two devices and estimates of MVPA change were not correlated regardless of the ActiGraph cut-points used. These results suggest that correlations between estimates of PA change are likely to be attenuated compared to the correlation between two PA measures during a single measurement period. This should not be surprising given that discrepancies in PA estimates at each measurement point will be compounded when quantifying the change in PA and that these additive errors will more easily mask the smaller within-subject changes in PA that are likely to occur over time. However, these findings should be interpreted with caution given the limited number of subjects with measures of PA change and the small magnitude of PA changes observed using either measure. Future studies should examine this issue in larger samples that have experienced clinically meaningful changes in PA.

A key strength of the current study is that it is one of the first studies to examine the validity of a wrist-worn Fitbit device in school-aged youth during free-living conditions. This addresses an important gap, as results from controlled laboratory studies may not reflect the magnitude of differences in device estimates in natural environments where children are likely to spend most of their time in non-structured physical activities. Prior studies have also been conducted in relatively small samples (19 to 59 children) and, to our knowledge, this is the first study to assess the ability of a wrist-worn Fitbit device to capture changes in youth physical activity over time. Finally, few prior studies in youth have examined the extent to which cut-point selection may influence the agreement between Fitbit and ActiGraph intensity-specific PA estimates.

This study also has several limitations and delimitations that should be considered when interpreting these results. We compared Flex 2 and ActiGraph estimates during concordant wear periods using an algorithm developed for use with ActiGraph count data but modified it to use minute by minute step data from both devices. Some misclassification of wear time is inevitable using this algorithm, and this misclassification may have been greater in the current study when applied to step data. The inclusion of minutes of non-concordant wear would increase discrepancies between device estimates and reduce the correlation between Fitbit and ActiGraph estimates of PA change. In addition, the differences observed between Flex 2 and ActiGraph estimates should not be attributed solely to error in the Fitbit device as the ActiGraph is not a gold standard measure and has known sources of error in measuring free-living PA in youth and adults (Dowd et al., 2018; Lynch et al., 2019). These potential errors in ActiGraph estimates include those attributable to the cut-points used to classify PA intensity, as clearly demonstrated by the current results. An additional source of error in our ActiGraph estimates is the use of intensity cut-points that were developed from data collected in 15-s epochs using earlier ActiGraph models. While studies have generally found good to strong agreement in counts and intensity classification across newer ActiGraph models (Clevenger et al., 2020; Montoye et al., 2018), the use of a different ActiGraph model may have introduced additional error in our intensity estimates. Lastly, commercial devices are constantly evolving with updates to firmware and the release of newer models with modified, but proprietary, algorithms (e.g., the Fitbit Ace, a newly introduced model targeted specifically for youth). Therefore, these results may not accurately reflect the level of accuracy of the same model at a later date or with subsequent models. The above issues likely introduce random error into both measures of physical activity, thereby increasing the magnitude of observed differences. Therefore, these results may underestimate the true level of agreement between the two devices.

In conclusion, the level of agreement between Fitbit Flex 2 and ActiGraph estimates in free-living youth may depend, in large part, upon the ActiGraph cut-points used to classify PA intensity. However, there is fair to good agreement between the Flex 2 and ActiGraph in ranking children's daily steps and MVPA regardless of the cut-points used, indicating the potential utility of wrist-worn wearables as a relative measure of free-living PA in youth. There may be lower concordance between these devices in assessing changes in PA over time. Future studies should examine the agreement between Fitbit and ActiGraph estimates of PA change in larger samples of free-living youth experiencing clinically meaningful levels of PA change.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Research reported in this publication was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number 1R01HL135359-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health; NHLBI.

ORCID

Michael D Schmidt http://orcid.org/0000-0002-8062-4491
Stephen L. Rathbun http://orcid.org/0000-0002-4704-5316
Lindsay Hahn http://orcid.org/0000-0002-0039-9782

References

Barlow, S. E., & Expert, C. (2007). Expert committee recommendations regarding the prevention, assessment, and treatment of child and adolescent overweight and obesity: Summary report. *Pediatrics*, *120*(4), S164–192. https://doi.org/10.1542/peds.2007-2329C

Buchele Harris, H., & Chen, W. (2018). Technology-enhanced classroom activity breaks impacting children's physical activity and fitness. *Journal of Clinical Medicine*, *7*(7), 165. https://doi.org/10.3390/jcm7070165

Byun, W., Kim, Y., & Brusseau, T. A. (2018). The use of a fitbit device for assessing physical activity and sedentary behavior in preschoolers. *The Journal of Pediatrics*, *199*, 35–40. https://doi.org/10.1016/j.jpeds.2018.03.057

Choi, L., Liu, Z., Matthews, C. E., & Buchowski, M. S. (2011). Validation of accelerometer wear and nonwear time classification algorithm. *Medicine & Science in Sports & Exercise*, *43*(2), 357–364. https://doi.org/10.1249/MSS.0b013e3181ed61a3

Clevenger, K. A., Molesky, M. J., Vusich, J., & Montoye, A. H. K. (2019). Free-living comparison of physical activity and sleep data from fitbit activity trackers worn on the dominant and nondominant wrists. *Measurement in Physical Education and Exercise Science*, *23*(2), 194–204. https://doi.org/10.1080/1091367X.2019.1577737

Clevenger, K. A., Pfeiffer, K. A., & Montoye, A. H. K. (2020). Cross-generational comparability of raw and count-based metrics from actigraph GT9X and wGT3X-BT accelerometers during free-living in youth. *Measurement in Physical Education and Exercise Science*, *24*(3), 194–204. https://doi.org/10.1080/1091367X.2020.1773827

Collins, J. E., Yang, H. Y., Trentadue, T. P., Gong, Y., Losina, E., & Parmenter, B. (2019). Validation of the fitbit charge 2 compared to the ActiGraph GT3X+ in older adults with knee osteoarthritis in free-living conditions. *PLoS One*, *14*(1), e0211231. https://doi.org/10.1371/journal.pone.0211231

Cradock, A. L., Barrett, J. L., Taveras, E. M., Peabody, S., Flax, C. N., Giles, C. M., & Gortmaker, S. L. (2019). Effects of a before-school program on student physical activity levels. *Preventive Medicine Reports*, *15*, 100940. https://doi.org/10.1016/j.pmedr.2019.100940

Dixon, P. M., Saint-Maurice, P. F., Kim, Y., Hibbing, P., Bai, Y., & Welk, G. J. (2018). A primer on the use of equivalence testing for evaluating measurement agreement. *Medicine & Science in Sports & Exercise*, *50*(4), 837–845. https://doi.org/10.1249/MSS.0000000000001481

Dominick, G. M., Winfree, K. N., Pohlig, R. T., & Papas, M. A. (2016). Physical activity assessment between consumer- and research-grade accelerometers: A comparative study in free-living conditions. *JMIR mHealth and uHealth*, *4*(3), e110–e110. https://doi.org/10.2196/mhealth.6281

Dowd, K. P., Szeklicki, R., Minetto, M. A., Murphy, M. H., Polito, A., Ghigo, E., van der Ploeg, H., Ekelund, U., Maciaszek, J., Stemplewski, R., Tomczak, M., & Donnelly, A. E. (2018). A systematic literature review of reviews on techniques for physical activity measurement in adults: A DEDIPAC study. *The International Journal of Behavioral Nutrition and Physical Activity*, *15*(1), 15. https://doi.org/10.1186/s12966-017-0636-2

Evans, E. W., Abrantes, A. M., Chen, E., & Jelalian, E. (2017). Using novel technology within a school-based setting to increase physical activity: A pilot study in school-age children from a low-income, Urban community. *BioMed Research International*, *2017*, 4271483. https://doi.org/10.1155/2017/4271483

Evenson, K. R., Catellier, D. J., Gill, K., Ondrak, K. S., & McMurray, R. G. (2008). Calibration of two objective measures of physical activity for children. *Journal of Sports Sciences*, *26*(14), 1557–1565. https://doi.org/10.1080/02640410802334196

Evenson, K. R., Goto, M. M., & Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *International Journal of Behavioral Nutrition and Physical Activity*, *12*(1), 159. https://doi.org/10.1186/s12966-015-0314-1

Feehan, L. M., Geldman, J., Sayre, E. C., Park, C., Ezzat, A. M., Yoo, J. Y., Hamilton, C. B., & Li, L. C. (2018). Accuracy of fitbit devices: Systematic review and narrative syntheses of quantitative data. *JMIR mHealth and uHealth*, *6*(8), e10527. https://doi.org/10.2196/10527

Fore, R., Hart, J. E., Choirat, C., Thompson, J. W., Lynch, K., Laden, F., Chavarro, J. E., & James, P. (2020). Embedding mobile health technology into the nurses' health study 3 to study behavioral risk factors for cancer. *Cancer Epidemiology, Biomarkers & Prevention*, *29*(4), 736–743. https://doi.org/10.1158/1055-9965.EPI-19-1386

Godino, J. G., Wing, D., de Zambotti, M., Baker, F. C., Bagot, K., Inkelis, S., Pautz, C., Higgins, M., Nichols, J., Brumback, T., Chevance, G., Colrain, I. M., Patrick, K., & Tapert, S. F. (2020). Performance of a commercial multi-sensor wearable (fitbit charge HR) in measuring physical activity and sleep in healthy children. *PLoS One*, *15*(9), e0237719. https://doi.org/10.1371/journal.pone.0237719

Gorzelitz, J., Farber, C., Gangnon, R., & Cadmus-Bertram, L. (2020). Accuracy of wearable trackers for measuring moderate- to vigorous-intensity physical activity: A systematic review and meta-analysis. *Journal for the Measurement of Physical Behaviour*, *3*(4), 346–357. https://doi.org/10.1123/jmpb.2019-0072

Hahn, L., Schmidt, M. D., Rathbun, S. L., Johnsen, K., Annesi, J. J., & Ahn, S. J. (2020). Using virtual agents to increase physical activity in young children with the virtual fitness buddy ecosystem: Study protocol for a cluster randomized trial. *Contemporary Clinical Trials*, *99*, 106181. https://doi.org/10.1016/j.cct.2020.106181. https://doi.org/ARTN 10618110.1016/j.cct.2020.106181.

Hao, Y., Ma, X. K., Zhu, Z., & Cao, Z. B. (2021). Validity of wrist-wearable activity devices for estimating physical activity in adolescents: Comparative study. *JMIR mHealth and uHealth*, *9*(1), e18320. https://doi.org/10.2196/18320

Hayes, L. B., & Van Camp, C. M. (2015). Increasing physical activity of children during school recess. *Journal of Applied Behavior Analysis*, *48*(3), 690–695. https://doi.org/10.1002/jaba.222

Kang, S., Kim, Y., Byun, W., Suk, J., & Lee, J. M. (2019). Comparison of a wearable tracker with actigraph for classifying physical activity intensity and heart rate in children. *International Journal of Environmental Research and Public Health*, *16*(15), 2663. https://doi.org/10.3390/ijerph16152663

Limb, E. S., Ahmad, S., Cook, D. G., Kerry, S. M., Ekelund, U., Whincup, P. H., Victor, C. R., Iliffe, S., Ussher, M., Fox-Rushby, J., Furness, C., Ibison, J., DeWilde, S., & Harris, T. (2019). Measuring change in trials of physical activity interventions: A comparison of self-report questionnaire and accelerometry within the PACE-UP trial. *International Journal of Behavioral Nutrition and Physical Activity*, *16*(1), 10. https://doi.org/10.1186/s12966-018-0762-5

Lynch, B. A., Kaufman, T. K., Rajjo, T. I., Mohammed, K., Kumar, S., Murad, M. H., Gentile, N. E., Koepp, G. A., McCrady-Spitzer, S. K., & Levine, J. A. (2019). Accuracy of accelerometers for measuring physical activity and levels of sedentary behavior in children: A systematic review. *Journal of Primary Care & Community Health*, *10*, 2150132719874252. https://doi.org/10.1177/2150132719874252

McLellan, G., Arthur, R., & Buchan, D. S. (2018). Wear compliance, sedentary behaviour and activity in free-living children from Hip-and wrist-mounted ActiGraph GT3X+ accelerometers. *Journal of Sports Sciences*, *36*(21), 2424–2430. https://doi.org/10.1080/02640414.2018.1461322

Migueles, J. H., Cadenas-Sanchez, C., Ekelund, U., Delisle Nystrom, C., Mora-Gonzalez, J., Lof, M., Labayen, I., Ruiz, J. R., & Ortega, F. B. (2017). Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports Medicine*, *47*(9), 1821–1845. https://doi.org/10.1007/s40279-017-0716-0

Montoye, A. H. K., Nelson, M. B., Bock, J. M., Imboden, M. T., Kaminsky, L. A., Mackintosh, K. A., McNarry, M. A., & Pfeiffer, K. A. (2018). Raw and count data comparability of hip-worn actigraph GT3X+ and link accelerometers. *Medicine & Science in Sports & Exercise*, *50*(5), 1103–1112. https://doi.org/10.1249/MSS.0000000000001534

Nicaise, V., Crespo, N. C., & Marshall, S. (2014). Agreement between the IPAQ and accelerometer for detecting intervention-related changes in physical activity in a sample of Latina women. *Journal of Physical Activity and Health*, *11*(4), 846–852. https://doi.org/10.1123/jpah.2011-0412

Plasqui, G., Bonomi, A. G., & Westerterp, K. R. (2013). Daily physical activity assessment with accelerometers: New insights and validation studies. *Obesity Reviews*, *14*(6), 451–462. https://doi.org/10.1111/obr.12021

Plasqui, G., & Westerterp, K. R. (2007). Physical activity assessment with accelerometers: An evaluation against doubly labeled water. *Obesity (Silver Spring)*, *15*(10), 2371–2379. https://doi.org/10.1038/oby.2007.281

Romanzini, M., Petroski, E. L., Ohara, D., Dourado, A. C., & Reichert, F. F. (2014). Calibration of ActiGraph GT3X, Actical and RT3 accelerometers in adolescents. *European Journal of Sport Science*, *14*(1), 91–99. https://doi.org/10.1080/17461391.2012.732614

Trost, S. G., McIver, K. L., & Pate, R. R. (2005). Conducting accelerometer-based activity assessments in field-based research. *Medicine & Science in Sports & Exercise*, *37*(11 Suppl), S531–543. https://doi.org/10.1249/01.mss.0000185657.86065.98