Physical Therapy   Pharmacy and Health Sciences, College of

2018

Current state of commercial wearable technology in physical activity monitoring 2015-2017

J. A. Bunn

J. W. Navalta

C. J. Fountaine

J. D. Reece

Current State of Commercial Wearable Technology in Physical Activity Monitoring 2015-2017

JENNIFER A. BUNN[1], JAMES W. NAVALTA[2], CHARLES J. FOUNTAINE[3], and JOEL D. REECE[4]

[1]College of Pharmacy and Health Sciences, Campbell University, Buies Creek, NC, USA; [2]Department of Kinesiology and Nutrition Sciences, University of Nevada, Las Vegas, Las Vegas, NV, USA; [3]Department of Applied Human Sciences, University of Minnesota Duluth, Duluth, MN, USA; [4]Department of Exercise Science, Brigham Young University-Hawaii, Laie, HI, USA

ABSTRACT

International Journal of Exercise Science 11(7): 503-515, 2018. Wearable physical activity trackers are a popular and useful method to collect biometric information at rest and during exercise. The purpose of this systematic review was to summarize recent findings of wearable devices for biometric information related to steps, heart rate, and caloric expenditure for several devices that hold a large portion of the market share. Searches were conducted in both PubMed and SPORTdiscus. Filters included: humans, within the last 5 years, English, full-text, and adult 19+ years. Manuscripts were retained if they included an exercise component of 5-min or greater and had 20 or more participants. A total of 10 articles were retained for this review. Overall, wearable devices tend to underestimate energy expenditure compared to criterion laboratory measures, however at higher intensities of activity energy expenditure is underestimated. All wrist and forearm devices had a tendency to underestimate heart rate, and this error was generally greater at higher exercise intensities and those that included greater arm movement. Heart rate measurement was also typically better at rest and while exercising on a cycle ergometer compared to exercise on a treadmill or elliptical machine. Step count was underestimated at slower walking speeds and in free-living conditions, but improved accuracy at faster speeds. The majority of the studies reviewed in the present manuscript employed different methods to assess validity and reliability of wearable technology, making it difficult to compare devices. Standardized protocols would provide guidance for researchers to evaluate research-grade devices as well as commercial devices used by the lay public.

KEY WORDS: Systematic review, wearable devices, exercise and fitness trackers, energy expenditure estimation, heart rate, step count, validity

INTRODUCTION

Commercial wearable technology has seen an expansive increase in personal use as well as application in industries including medicine, healthcare, and the military. Within the area of health and fitness, wearable technology was determined to be the top worldwide fitness trend in 2016 (14), and 2017 (15). Because of this growing trend, and the ubiquity of smart-linked

apparatuses, it becomes important to determine the validity and reliability of wearable devices that are available to the general public.

Unfortunately, there appears to be an evidence-based research lag with respect to reporting the accuracy of commercially available devices. In some cases, by the time a study has gained approval, participants have been tested, data analyzed, and reports have been written and gone through the peer review process, a wearable technology device has been updated to the next model or has become obsolete (10). Thus, one purpose of this article is to provide a current systematic review of recent high-quality literature relating to wearable technology devices. It is hoped that the synthesis of this data can aid researchers desiring to utilize a device for a particular application in the selection of the most appropriate item.

Additionally, there is often confusion on the part of consumers regarding which device is optimal for personal use. It is hoped that the results of this systematic review can assist consumers to make informed decisions when deciding to purchase a particular device. Toward this end, we have included evidence-based summaries of specific devices. We hope that this paper can be a resource for both researchers as well as personal consumers wishing to utilize wearable technology devices for physical activity monitoring.

METHODS

Protocol
The most popular devices on the market (6) were chosen for this systematic review (12) and included: Fitbit, Garmin, Apple, Misfit, Samsung Gear, TomTom, and Lumo. The searched terms were: "(Device name) AND (validity OR validation OR validate OR comparison OR comparisons OR comparative OR reliability OR accuracy). Searches were conducted in both PubMed and SPORTdiscus. Search filters included: humans, last 5 years, English, full-text, and adult 19+ years. One researcher went through the articles listed for each device and determined if the manuscript was truly related to validation, accuracy, and reliability. Only articles related to these topics were deemed pertinent and were retained for further assessment. Table 1 shows the number of articles found and relevant for each device searched. Of these articles, those that had an exercise component of 5-min in duration or greater and had 20 participants or more were included in order to correspond with the Consumer Technology Association standard for measuring step counts on consumer wearable activity monitoring devices (4). For energy expenditure assessment, only studies that used a validated metabolic analysis system as the criterion measure were included. For step count, only studies using video or visual step count as the criterion measure were included. Studies that solely assessed sleep or exercise rehabilitation and those that were included in previous systematic reviews (7) were excluded. Two independent evaluators utilized the stated criteria, and agreed upon the final papers that were included, N = 10. Several articles included assessment of multiple devices, including devices that were not part of the original search. Data on these devices are reported in this systematic review.

Table 1. Number of articles found in the original search and after assessment for relevance to wearable technology, validity, accuracy, and reliability.

Device	Articles found	Pertinent articles
Fitbit	41	24
Garmin	11	11
Apple	106	9
Misfit	68	5
Samsung Gear	2	2
TomTom	4	2
Lumo	0	0

ENERGY EXPENDITURE

In general, the Fitbit family of products seem to have the higher validity compared to other wearable devices when estimating energy expenditure, and Jawbone lower (see table 2). Overall, wearable devices tend to underestimate energy expenditure compared to criterion laboratory measures (Oxycon Mobile, CosMed K4b2, or MetaMax 3B), however at higher intensities of activity energy expenditure is underestimated. Additionally, while wearable technology devices are better at estimating energy expenditure during sedentary activities, validity becomes poorer as exercise begins, and gets worse as the intensity increases. Future studies should continue to determine the optimal levels of intensity that will return acceptable validity measurements. Regarding the analysis of validity, all studies reviewed utilized the Bland-Altman procedure for determining agreement (3, 13, 16, 17), two of four incorporated correlation techniques (Pearson Product Moment, or Rho) (3, 16), and three of four reported the mean absolute percent error (3, 13, 17). It is recommended that future investigations utilize all three indicators of validity to allow for ease of comparison between devices. Furthermore, it is recommended that a common unit of measure be reported with respect to energy expenditure to allow comparison between studies. In this case, three of four studies reported energy expenditure in terms of overall calories expended (kcal). No study included in this current systematic review of energy expenditure estimation in wearable technology devices reported test-retest reliability measures. Future studies should include this measure so that consumers and researchers will be able to make informed decisions regarding both validity and reliability of devices.

Table 2. Summary table of current investigations determining validity of various wearable technology devices compared to a criterion measure for energy expenditure.

Reference	Subjects	Activity	Validity – Agreement (95% confidence interval range)	Validity – Correlation (r-values)	Validity – MAPE (%)
Chowdhury et al. 2017 (3)	N = 30 (15 male, 15 female) 27±6 yrs	24-min activities of daily living 64-min exercise (10-min each: treadmill walking, walking with bags, cycling, treadmill running)	Apple Watch (0.2±3.4) > Fitbit Charge HR (0.3±4.6) > Microsoft Band (-1.8±3.9) > Jawbone UP24 (-0.9±5.4) kcal/min	Apple Watch (0.935) > Microsoft Band (0.879) > Fitbit Charge HR (0.825) > Jawbone UP24 (0.800)	Apple Watch (27±19%) < Jawbone UP24 (36±14%) < Fitbit Charge HR (36±22%) < Microsoft Band (40±16%)
Nelson et al. 2016 (13)	N = 30 (15 male, 15 female) 10 each in age groups 18-39 yr, 40-59 yr, 60-80 yr	5-min ambulatory/ exercise of increasing intensity including: walking in hallway, treadmill walking, cycling, hallway jog, treadmill jog	Fitbit One (159.0 to 127.4) > Fitbit Flex (180.7 to 147.0) > Fitbit Zip (189.4 to 155.1) > Jawbone UP (162.6 to 127.8) kcal	-	WALKING: Jawbone UP (24%) < Fitbit One (31%) < Fitbit Flex (53%) < Fitbit Zip (68%) JOGGING: Fitbit One (22%) < Fitbit Flex (35%) < Fitbit Zip (37%) < Jawbone UP (46%)
Wallen et al. 2016 (16)	N = 22 (11 male, 11 female) 24.0±5.6 yrs	5-min sedentary, 3-min stages walking, 3-min stages cycling	Samsung Gear S (-73.5 to 21.3) > Fitbit Charge HR (-137.0 to 17.3) > Apple Watch (-219.7 to -12.9) > Mio ALPHA (-266.7 to 65.7) kcal	Samsung Gear S (0.86) > Fitbit Charge HR (0.64) > Mio ALPHA (0.46) > Apple Watch (0.16)	-
Woodman et al. 2017 (17)	N = 28 (20 male, 8 female) 25.5±3.7 yrs	10-min sedentary, 5-min activities increasing intensity: treadmill walking, overground walking, overground running, overground cycling, laboratory cycling	Garmin VivoFit (93.8 to 271.8) > Withings Hip (56.7 to 282.8) > Withings Shirt (59.8 to 286.2) > Withings Wrist (142.7 to 382.6) > Basis Peak (-290.4 to 233.1) kcal	-	Basis Peak (27.2%) < Withings Pulse Hip (40.3%) < Withings Pulse Shirt (41.4%) < Garmin VivoFit (44.6%) < Withings Pulse Wrist (63.7%)

HEART RATE

Table 3. Summary table of current investigations determining validity of various wearable technology devices compared to a criterion measure for heart rate.

Reference	Subjects	Activity	Criterion	Validity - Correlation	Validity - % Difference
Gillinov et al. 2017 (9)	N = 50 (27 females, 23 males); 38±12 years	4.5 min at 3 intensities on each piece of equipment: treadmill, cycling, elliptical w/ and w/o arms	12-lead ECG	Polar H7 (0.99) > Apple Watch (0.92) > TomTom Spark (0.83) > Garmin 235 (0.81) > Scosche Rhythm (0.75) > Fitbit Blaze (0.75)	MAPE reported for each specific exercise, but the order typically was Polar < Apple Watch < TomTom Spark < Scosche Rhythm < Garmin 235 < Fitbit Blaze
Wallen et al. 2016 (16)	N = 22 (11 females, and 11 males), 24.9±5.6 years	1-hr involving rest, treadmill walking and running, and cycling	3-lead ECG	Apple Watch (0.98) > Mio Alpha (0.91) > Samsung Gear (0.80) > Fitbit Charge HR (0.78)	% Difference: Apple Watch (-1.3±4.4) < Mio Alpha (-4.3±7.2) < Samsung Gear (-7.1±10.3) < Fitbit Charge HR (-9.3±8.5)
Jo et al. 2016 (11)	N = 24 (12 females, 12 males), 24.8±2.1 years	77-min protocol involving rest, treadmill walking and running, cycling at 2 different intensities, and strength training	12-lead ECG	Basis Pak (.92) > Fitbit Charge HR (.83)	MAPE: Basis Peak (5.3±8.3) < Fitbit Charge HR (9.8±14.0)

Each of the studies reviewed included an analysis of error (either mean percent error or absolute percent error), a correlation assessment with the criterion (either Lin's concordance, intraclass correlations, or Pearson Product Moment), and utilized the Bland-Altman method for evaluating agreement and error. Criterion assessment in the studies evaluated included an electrocardiogram (ECG) or Polar chest strap. Wrist and forearm activity monitors had a wide range of accuracy, with the Apple iWatch having the lowest mean absolute percent error (MAPE) and the Fitbit devices having the highest MAPE. The details of the study are shown in Table 3. All wrist and forearm devices had a tendency to underestimate heart rate, and this error was generally greater at higher exercise intensities and those that included greater arm movement. Heart rate measurement was also typically better at rest and while exercising on a cycle ergometer compared to exercise on a treadmill or elliptical machine. One study included a Polar chest strap as a tested device compared to an ECG, and the chest strap had the lowest MAPE and highest concordance compared to the wrist and forearm devices.

The three studies assessed used different correlation assessments and methods for evaluating error. An industry standard for reporting these two values would be useful. Of equal importance is to ensure proper wear of the devices and avoid simultaneously wearing multiple devices on one arm. Devices were worn properly and according to manufacturers'

guidelines in all three of the heart rate studies assessed in this review. Only one of the studies used continuous heart rate assessment (recorded by the devices each second) and the other two studies recorded heart rate at specific times after reaching steady state during exercise. Only the Fitbit Charge HR device was tested in both types of studies, and the results of the continuous study were less favorable. Assessing heart rate at specific intervals after reaching steady state eliminates the oscillation of heart rate with changing intensities. The response rate of the device is important to assess with changing intensities and second-by-second analysis for heart rate accuracy should be encouraged in development and evaluation of these devices. Further, this is important because devices have shown to have lag in readings of heart rate and data dropout. Gillinov et al. was the only study to address this issue, and the authors were transparent regarding which devices had errors and how many data points were missed or removed (9). While this transparency is useful, it falsely increases the accuracy of the devices by removing bad data.

STEP COUNT

Each of the studies reviewed utilized either a direct hand-tally count (1, 8, 13) or video recording (2, 10) to serve as the criterion measure for step count. The methods used to assess validity were varied amongst the studies. Three of the five studies used MAPE (1, 8, 13), whereas two of the five used absolute percentage error (APE) (2, 10). Three of the five studies (1, 2, 8) utilized Bland-Altman plots to show 95% limits of agreement, but only one study (1) calculated correlation coefficients. All five studies utilized a treadmill protocol, with speeds ranging from 2-5 mph, while spending 3-10 minutes at each incremental speed. In addition to a treadmill protocol, three of the five studies also included an over-ground condition (1, 10, 13), and two studies (1, 13) examined validity in free-living conditions.

Collectively, there was wide variability and accuracy across the various physical activity monitor brands for both speed and condition, as shown in Table 4. The studies reviewed consistently demonstrated reduced validity in terms of underestimating step counts at both slower walking speeds (< 2 mph), ambulatory, and free-living conditions, but improved accuracy at faster speeds, which is consistent with previous research. The Fitbit One and Fitbit Zip consistently demonstrated MAPE < 5% and were noted by multiple studies to be the most accurate (1, 10, 13). Conversely, multiple studies (1, 8, 10) found the Nike+ FuelBand and Polar Loop to be the least accurate, with MAPE >10%.

Challenges for future research of the step count feature on activity trackers are not unlike what was observed within energy expenditure and heart rate validation studies. Whereas the treadmill protocols reviewed adhered relatively close to Consumer Technology Association (CTA) step count standards (4), over-ground and free-living conditions lack a standardized protocol, and are a notable limitation within the literature. In regards to the assessment of validity, no standardized threshold exists for what constitutes high or low MAPE, thus in the studies reviewed, a wide range of cutoff criterion for acceptable MAPE was observed. Likewise, whereas Bland-Altman plots are commonly constructed to show limits of agreement, the studies reviewed were very inconsistent in actually providing 95% limits of agreement

between the criterion measure and each respective device (mean difference ± 1.96 SD of the differences), making direct comparisons difficult.

Table 4. Summary table of current investigations determining validity of various wearable technology devices compared to a criterion measure for heart rate.

Reference	Subjects	Activity	Device(s)	Validity - MAPE	Validity – correlation to criterion
An et al. 2017 (1)	N = 35 (18 females, 17 males), 31.0±11.8 yrs	Treadmill: 3 minutes at 3.2, 4, 4.8, 5.6, 6.4, 8.04 km/h Over-ground: indoor track 24 hr Free-living	Fitbit Zip, Withings Pulse, Jawbone UP 24, Basis B1 Band, Garmin VivoFit, Sense Wear Mini, Fitbit Flex, Misfit Shine, Polar Loop, Nike + FuelBand	All devices and treadmill speeds = 8.2% All devices over-ground = 9.9% Free-living = 18.48%	Fitbit Zip & Withings Pulse r=1.0 > Jawbone UP24 & SenseWear Mini r=0.9 > Basis B1 Band, Fitbit Flex, Misfit Shine, & Nike+Fuel Band r=0.8 > Garmin VivoFit & Polar Loop r=0.7
Chen et al. 2016 (2)	N = 30 (15 females, 15 males), 21.5±2.0 yrs	Treadmill: 5 minutes at 3.2, 4.8, 6.4, 8.04 km/h Other: 6 Simulated daily activities	Fitbit Flex, Garmin VivoFit, Jawbone UP,	Absolute Percent Error (APE) all devices and speeds = 1.5-9.6% APE for 8.04 km/h with all devices < 2.5%	Devices over counted steps At 4.8 km/h: JawboneUP = 64.4 Gamin Vivofit = 87.8 Fitbit Flex = 157 And at 8.04 km/h: JawboneUP = 54.6 Gamin Vivofit = 85.4 Fitbit Flex = 79.1
Fokkema et al. 2017 (8)	N = 31 (15 females, 16 males), 32.0±12.0	Treadmill: 10 minutes at 3.2, 4.8, 6.4 km/h	Garmin VivoSmart, Fitbit Charge HR, Polar Loop, Apple Watch Sport, Pebble Smartwatch, Samsung Gear S, Misfit Flash, Jawbone UP Move, Flyfit, Moves	All devices = 0.0-26.4% Best Devices: Garmin VivoSmart = -0.2-9.0% Fitbit Charge HR = -.07-5.2% Apple Watch Sport = 0.0-1.9%	All devices ICC = -0.02-0.97, Slow 3.2 km/hr ICC = 0-0.95 Average 4.8 km/hr ICC = 0-0.98 Vigorous 6.4 km/hr ICC = 0-0.92

Huang et al. 2016 (10)	N = 40 (15 females, 25 males), 23.9±2.8 yrs	Treadmill: 3 minutes at 3.2, 4.8, 6.4 km/h Flat ground testing Stairs testing	Nike+FuelBand SE, Jawbone UP 24, Fitbit One, Fitbit Flex, Fitbit Zip, Garmin Vivofit, Yamax CW-701, Omron HJ-321	All devices = 0.1-16.7% At level walking all <1% except Garmin Vivofit, Fitbit Flex, & Nike+FuelBand SE All devices on stairs = 1.1-7.9%, except Nike+FuelBand SE = 34.3%	Trends of systematic bias for Jawbone UP 24 (slope = 0.4, R = 0.20), Garmin Vivofit (slope = 0.4, R = 0.19), Fitbit Flex (slope = 0.8, R = 0.49) and Nike+FuelBand SE (slope = 1.1, R = 0.54)
Nelson et al. 2016 (13)	N = 30 (15 females, 15 males), 48.9±19.4 yrs	5 minutes of self-paced sedentary, household, ambulatory, walking, jogging, stairs, cycling	Fitbit One, Fitbit Flex, Fitbit Zip, Jawbone UP 24, Omron HJ-720IT (criterion)	All monitors for Household activities = 54-79%, Ambulatory activities= 3-6%, Walking = 2-11%, Jogging = 3-8%, Cycling = 70-93%	Sedentary: No monitors significantly differed from researcher step count, Household: Only Fitbit One not statistically different from Omron, Ambulatory: FitbitZip significantly more steps than Omron, Walking, Jogging, Stairs: no statistical difference from Omron except FitbitFlex Cycling: Only Jawbone significantly different from Omron

WEARABLE TECHNOLOGY DEVICES

Apple Watch
The Apple Watch has been evaluated in four recent investigations (3, 8, 9, 16). Evidence indicates that validity to criterion measures is acceptable for heart rate and step count (see table 5). Researchers and consumers should view energy expenditure output with caution.

Table 5. Average validity measurements (Intraclass correlation, ICC; Mean average percentage error, MAPE; and Agreement) for energy expenditure, heart rate, and step count in the Apple watch.

	Energy Expenditure			Heart Rate			Step Count		
	ICC	MAPE	Agreement	ICC	MAPE	Agreement	ICC	MAPE	Agreement
Apple Watch	0.493	27%	-232 to -14	0.95	-	-13.5 to 14.6	0.727	1.08%	-69.8 to 92.0

Basis Band (B1, Peak)

Two devices from the Basis brand were evaluated in recent studies (1, 11, 17). The devices appear to be valid for steps and heart rate, but not for energy expenditure (see Table 6). Both of these devices have been discontinued.

Table 6. Average validity measurements (Intraclass correlation, ICC; Mean average percentage error, MAPE; and Agreement) for energy expenditure, heart rate, and step count in Basis discontinued devices.

	Energy Expenditure			Heart Rate			Step Count		
	ICC	MAPE	Agreement	ICC	MAPE	Agreement	ICC	MAPE	Agreement
B1	0.138	23.5%	-78.9 to 248.3	-	-	-	0.7	3.1%	-63 to 100.1
Peak	0.022	27.2%	-	0.935	4.5	-24 to 79.9	-	-	-

Fitbit Family of Devices

The Fitbit Charge HR had the greatest influence in the recent literature (3, 8, 11, 16) and has good validity for heart rate (see table 7). Validity for energy expenditure and step count are lower than what is observed for heart rate.

The Fitbit Flex was utilized in three recent investigations that evaluated step validity (1, 2, 13). All studies reported MAPE, and taken together are outside of the acceptable 5-10% error for controlled or free-living investigations. However, one study reported high ICC and relatively narrow limits of agreement (1). Only one study using the Fitbit Flex determined that energy expenditure MAPE was greater than acceptable error (13).

Similarly, the Fitbit Zip was observed to have higher than acceptable step MAPE in two studies (1, 13), but a good ICC and narrow limits of agreement in the only investigation to report these values (1). Again, only one study using the Fitbit Zip determined that energy expenditure MAPE displayed greater than acceptable error (13).

A single recent investigation has evaluated the Fitbit One, and found poor MAPE values for both step and energy expenditure (13). Finally, a single investigation utilizing the Fitbit Blaze found that heart rate was valid (9).

Table 7. Average validity measurements (Intraclass correlation, ICC; Mean average percentage error, MAPE; and Agreement) for energy expenditure, heart rate, and step count in Fitbit devices.

	Energy Expenditure			Heart Rate			Step Count		
	ICC	MAPE	Agreement	ICC	MAPE	Agreement	ICC	MAPE	Agreement
Charge HR	0.693	36%	-137 to 17.3	0.805	-	-34 to 23	0.526	3.03%	-108 to 70.5
Flex	-	34%	-	-	-	-	0.80	14.56%	-41.1 to 101.8
Zip	-	39.8%	-	-	-	-	1.0	22.18%	-8.7 to 10.1
One	-	25.4%	-	-	-	-	-	25%	-
Blaze	-		-	0.67	-	-30 to 45	-	-	-

Garmin Family of Devices
The Garmin Vivofit was evaluated in three recent studies (1, 2, 17). While this device appears to be valid for counting steps (see table 8), energy expenditure measurements are outside of acceptable limits. The Vivofit device itself does not provide a measure of heart rate (it can be paired with a heart rate monitor to read through the device), and as such does not have an individual assessment for this variable.

A single recent investigation evaluated the Garmin Vivosmart (8), and focused exclusively on the step count measurement. Overall, the Vivosmart can return an accurate step count at most walking speeds (see table 8).

Only one recent study evaluated the Garmin Forerunner 235 (9), and opted to evaluate heart rate as the sole dependent variable. Overall, the Forerunner 235 provides valid heart rate measurements (see table 8).

Table 8. Average validity measurements (Intraclass correlation, ICC; Mean average percentage error, MAPE; and Agreement) for energy expenditure, heart rate, and step count in Garmin devices.

	Energy Expenditure			Heart Rate			Step Count		
	ICC	MAPE	Agreement	ICC	MAPE	Agreement	ICC	MAPE	Agreement
Vivofit	-	44.6%	-93.8 to 271.8	-	-	-	0.75	5.5%	-65.1 to 103.7
Vivosmart	-	-	-	-	-	-	0.592	3.9%	-89.3 to 183.3
Forerunner 235	-	-	-	0.81	-	-27 to 33	-	-	-

Jawbone Up24, Move
Jawbone devices were recently evaluated in four studies (1-3, 13) for steps and energy expenditure (table 9). Neither device measures heart rate. The Jawbone Move device only measures steps and sleep and can be placed either on the waist or worn around the wrist. The Move is more accurate for steps when worn at waist level. The Up24 device has been discontinued by Jawbone.

Table 9. Average validity measurements (Intraclass correlation, ICC; Mean average percentage error, MAPE; and Agreement) for energy expenditure, heart rate, and step count in Jawbone devices.

	Energy Expenditure			Step Count		
	ICC	MAPE	Agreement	ICC	MAPE	Agreement
Up24	0.77	33.3%	-101.3 to 147.5	0.75	14.1%	-36.9 to 47.5
Move	-	-	-	0.81	5.3%	-265 to 396

Misfit Family of Devices
One investigation evaluated the Misfit Shine (1), and another the Misfit Flash (8). While both devices provide an estimate of energy expenditure, the studies evaluated step count only. The accuracy of step count in these devices is low (see table 10).

Table 10. Average validity measurements (Intraclass correlation, ICC; Mean average percentage error, MAPE; and Agreement) for step count in Misfit devices.

	Energy Expenditure			Step Count		
	ICC	MAPE	Agreement	ICC	MAPE	Agreement
Shine	-	-	-	0.60	12.1%	-52.2 to 113.1
Flash	-	-	-	0.122	10.1%	-356.5 to 569

Polar Loop
Step count validity of the Polar Loop was evaluated in two recent studies (1, 8). The device can return an estimate of energy expenditure, and must be connected to a separate heart rate monitor, however these variables were not evaluated. Average validity measurements indicate that step count validity of the Polar Loop is low (ICC = 0.460, MAPE = 15.33%, Agreement = -161.4 to 328.1).

Samsung Gear S
Two recent studies reported validity measurements in the Samsung Gear S (8, 16). Energy expenditure estimates and heart rate appears to be valid (see table 11). Step count obtained from the Samsung Gear S has acceptable ICC and MAPE, but wide limits of agreement.

Table 11. Average validity measurements (Intraclass correlation, ICC; Mean average percentage error, MAPE; and Agreement) for energy expenditure, heart rate, and step count in the Samsung Gear S.

	Energy Expenditure			Heart Rate			Step Count		
	ICC	MAPE	Agreement	ICC	MAPE	Agreement	ICC	MAPE	Agreement
Samsung Gear S	0.86	-	-73.5 to 21.3	0.80	-	-27.3 to 13.1	0.605	3.3%	-204.7 to 223.3

Withings Pulse
One recent investigation found the Withings Pulse returned valid step count measurements (ICC = 0.95, MAPE = 1.65%, Agreement = -16.8 to 23.4) (1). However, a different study reported unacceptable energy expenditure validity (MAPE = 35%, Agreement = 86.4 to 317.2) (17).

RECOMMENDATIONS

The majority of the studies reviewed in the present manuscript employed different methods to assess validity and reliability of wearable technology. This difference in protocols makes it difficult to compare devices. The Consumer Technology Association (CTA) recently published validation criteria and protocols to evaluate devices in a standardized format (4). The CTA standard is set up for laboratory-based assessment of steps only, but provides a strong basis for comparison of devices. The CTA has also published a standard for evaluating devices for sleep validity (5), and a standard for heart rate is expected to be released in 2018. More standards and protocols should be developed to include heart rate, energy expenditure, and free-living conditions. These standards would provide guidance for researchers to evaluate research-grade devices as well as commercial devices used by the lay public.

Following these guidelines, it is recommended that exercise duration be at least 5-minutes in length in order to allow subjects to attain steady state measures. Additionally, regardless of exercise mode, it is suggested that at least two different exercise intensities be employed to allow for comparison. Furthermore, as there is a need to obtain reliability measures on devices, it is recommended that study designs be utilized to address this need. Thus future research on wearable technology devices should address both the question of validity as well as reliability. With respect to determining accuracy, it is recommended that future investigations address validity utilizing ICC, MAPE, and agreement to an established criterion measure (i.e. Bland-Altman analysis) in order to present overall evidence of validity. As the landscape of wearable technology devices is expanding, producing high-quality evidence of device accuracy and reliability will continue to be important to investigators wishing to utilize these items for research as well as general consumers that employ them for personal use.

REFERENCES

1. An HS, Jones GC, Kang SK, Welk GJ, Lee JM. How valid are wearable physical activity trackers for measuring steps? Eur J Sport Sci 17(3):360-368, 2017.

2. Chen MD, Kuo CC, Pellegrini CA, Hsu MJ. Accuracy of wristband activity monitors during ambulation and activities. Med Sci Sports Exerc 48(10):1942-1949, 2016.

3. Chowdhury EA, Western MJ, Nightingale TE, Peacock OJ, Thompson D. Assessment of laboratory and daily energy expenditure estimates from consumer multi-sensor physical activity monitors. PloS one 12(2):e0171720, 2017.

4. Consumer Technology Association. Physical activity monitoring for fitness wearables: Step counting. ANSI/CTA Standard 2016.

5. Consumer Technology Association. Methodology of measurements for features in sleep tracking consumer technology devices and applications. ANSI/CTA Standard 2017.

6. Duffy J, Colon A. The best fitness trackers of 2018. PC Reviews.

7. Evenson KR, Goto MM, Furberg RD. Systematic review of the validity and reliability of consumer-wearable activity trackers. Int J Behav Nutr Phys Activity 12:159, 2015.

8. Fokkema T, Kooiman TJ, Krijnen WP, CP VDS, M DEG. Reliability and validity of ten consumer activity trackers depend on walking speed. Med Sci Sports Exerc 49(4):793-800, 2017.

9. Gillinov S, Etiwy M, Wang R, Blackburn G, Phelan D, Gillinov AM, Houghtaling P, Javadikasgari H, Desai MY. Variable accuracy of wearable heart rate monitors during Aaerobic exercise. Med Sci Sports Exerc 49(8):1697-1703, 2017.

10. Huang Y, Xu J, Yu B, Shull PB. Validity of FitBit, Jawbone UP, Nike+ and other wearable devices for level and stair walking. Gait Posture 48:36-41, 2016.

11. Jo E, Lewis K, Directo D, Kim MJ, Dolezal BA. Validation of biofeedback wearables for photoplethysmographic heart rate tracking. J Sports Sci Med 15(3):540-7, 2016.

12. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 6(7):e1000100, 2009.

13. Nelson MB, Kaminsky LA, Dickin DC, Montoye AH. Validity of consumer-based physical activity monitors for specific activity types. Med Sci Sports Exerc 48(8):1619-1628, 2016.

14. Thompson WR. Worldwide survey of fitness trends for 2016. ACSMs Health Fit J 19(6):9-18, 2015.

15. Thompson WR. Worldwide survey of fitness trends for 2017. ACSMs Health Fit J 20(6):8-17, 2016.

16. Wallen MP, Gomersall SR, Keating SE, Wisloff U, Coombes JS. Accuracy of heart rate watches: implications for weight management. PloS one 11(5):e0154420, 2016.

17. Woodman JA, Crouter SE, Bassett DR, Jr., Fitzhugh EC, Boyer WR. Accuracy of consumer monitors for estimating energy expenditure and activity type. Med Sci Sports Exerc 49(2):371-377, 2017.