International Journal of EXERCISE SCIENCE

Original Research

Step Count Reliability and Validity of Five Wearable Technology Devices While Walking and Jogging in both a Free Motion Setting and on a Treadmill

JEFFREY MONTES†, RICHARD TANDY‡, JOHN YOUNG‡, SZU-PING LEE‡, and JAMES W. NAVALTA‡

Department of Kinesiology and Nutrition Sciences, University of Nevada, Las Vegas, Las Vegas, NV, USA

†Denotes graduate student author, ‡Denotes professional author

ABSTRACT

International Journal of Exercise Science 13(7): 410-426, 2020. Wearable technology devices are used by millions of people who use daily step counts to promote healthy lifestyles. However, the accuracy of many of these devices has not been determined. The purpose was to determine reliability and validity of the Samsung Gear 2, FitBit Surge, Polar A360, Garmin Vivosmart HR+, and the Leaf Health Tracker when walking and jogging in free motion and treadmill conditions. Forty volunteers completed walking and jogging free motion and treadmill protocols of 5-minute intervals. The devices were worn simultaneously in randomized configurations. The mean of two manual steps counters was used as the criterion measure. Test-retest reliability was determined via Intraclass Correlation Coefficient (ICC). Validity was determined via a combination of Pearson's Correlation Coefficient, mean absolute percent error (MAPE: free motion ≤ 10.0%, treadmill ≤ 5.00%), and Bland-Altman analysis (device bias and limits of agreement). Significance was set at $p < 0.05$. The Samsung Gear 2 was deemed to be both reliable and valid for the jogging conditions, but not walking. The Fitbit Surge was reliable and valid for all conditions except for treadmill walking (deemed reliable, ICC = 0.76; but not valid). The Polar A360 was found to be reliable for one condition (treadmill jog ICC = 0.78), but not valid for any condition. The Garmin Vivosmart HR+ and Leaf Health Tracker were found to be both reliable and valid for all situations. While each device returned some level of consistency and accuracy during either free motion or treadmill exercises, the Garmin Vivosmart HR+ and the Leaf Health Tracker were deemed to be reliable and valid for all conditions tested.

KEY WORDS: Step count, accuracy testing, inter-rater reliability, test-retest reliability, wearable technology

INTRODUCTION

Obesity rates in the United States are an important health issue. The Center for Disease Control and Prevention estimates that 39.8% of adults and 18.5% of youth are classified as obese with corresponding annual medical costs of $147 billion in 2008 US dollars (or $1,492 per person) (8). It projects that only 30.8% of the population is at a healthy recommended weight (8). However, because obesity has been linked to increased risks of cardiovascular disease, stroke, myocardial infarction, and diabetes, this yearly financial cost may actually be as high as $320.1 billion (27).

In order to combat this health affliction, reduce the associated financial burden, and promote healthy lifestyles, the Healthy People 2020 initiative has targeted a 3.1% population increase for those whose weight is to be within appropriate healthy recommendations (9). Achieving this goal requires various strategies to promote physical activity in the overweight/obese population to include cardiovascular, muscular, and daily activity movements to improve a daily healthy lifestyle.

A common objective for healthy living that is both easy to promote and understand is walking at least 10,000 steps every day. This idea of walking for health using a daily stepping goal has been employed for decades beginning with early pedometer manufacturers (3). Current research supports the monitoring of daily step counts and how it positively influences daily physical activity, health, and wellness levels (39). The American College of Sports Medicine (ACSM) recommends all persons do at least 30 minutes of moderate-intensity physical activity on at least 5 days a week (21). It has been estimated that the average U.S. adult takes approximately 6,500 steps per day, and that the ACSM's recommended daily activity requirement could be met by taking an additional 3,500 steps (11). Furthermore, scientific literature has provided evidence that taking 10,000 steps per day may allow for persons to "burn" up to 20% of their daily caloric requirement (18). However, while 10,000 steps a day has been shown to provide general health benefits, 15,000 steps a day may actually be necessary to decrease the risk of more serious conditions such as cardiovascular disease (38). Regardless, daily step counts can be viewed as a key component in maintaining health and helping prevent metabolic diseases.

Wearable technology has been rated the top fitness trend for the past two years (34, 36, 37) and based on forecasted financial trends, its use is expected to grow every year for the near future (35). Recent investigations have tested step count wearable technology in the laboratory (10, 17) and during flat ground walking and/or stair climbing (1, 19) with varying results of accuracy. However, none to our knowledge have evaluated the same wearable technology device in both a laboratory and free motion setting while performing basic movements such as walking and jogging. The common belief among researchers is that wearable technology is more accurate in a controlled setting such as on a treadmill (19). However, the need to evaluate the accuracy of these devices in both settings is important. While some people can exercise outside in a free motion setting, some prefer to be inside on a treadmill due to convenience, because of extreme outdoor weather conditions, or environmental concerns such as air pollution levels. Also, because of the proprietary algorithms used by each device to detect what criteria registers as a step, it is necessary to evaluate each with similar protocols in order to provide feedback as to whether the utilized measuring method is performing as expected in common situations.

Guidelines suggested by the Consumer Technology Association (CTA) for validating wearable technology step count measurements recommends that video recordings be made of any activity performed with two reviewers independently watching the video and producing identical manual step counts (12). In a free motion setting, this would be difficult and unfeasible in certain settings due to the potential for visual obstructions, interference from the public, or the lack of portable recording equipment. For example, a recent investigation reported step count validity

during the free motion activities of hiking and trail running (28). As the investigators were unable to implement of the CTA recommendations for video recording in this environment, two independent step counters returned high intra-rater reliability (ICC range = 0.991 to 0.996) (28). It is unknown how manual step counts taken by independent raters compare in a laboratory setting versus free motion conditions.

The purpose of this research is threefold: 1) to determine if the tested wearables are reliable for step count measurements when free motion walking, free motion jogging, treadmill walking, and treadmill jogging, 2) to determine if the devices would also be valid in the same conditions, and 3) to determine the intra-rater reliability of visual step counts by two independent counters. Based on our previous investigations utilizing wearable technology (23, 25, 28, 35), it was hypothesized that all five wearable technology devices would be reliable and valid under all four conditions. It was also hypothesized that manually obtained step counts from the two independent evaluators would return acceptable intra-rater reliability values.

METHODS

Participants

Forty healthy (identified as low risk according to the ACSM pre-participation screening questionnaire) (male n = 20, female n = 20) volunteered for this investigation with the following descriptive characteristics: age = 25 ± 7 years, height = 169.64 ± 11.18 cm, mass = 77.19 ± 19.2 kg, body mass index = 26.43 ± 5.19 m/kg^2. Participants filled out an informed consent form that was approved by the UNLV Biomedical Institutional Review Board (#885569-3). This research was carried out fully in accordance to the ethical standards of the International Journal of Exercise Science (29).

Devices: The Samsung Gear 2 (Samsung Electro-Mechanics, Seoul, South Korea) is a wrist-worn smartwatch. Sensors include an accelerometer, gyroscope, and heart rate monitor. The Fitbit Surge (Fitbit Inc, San Francisco, CA) is a fitness super wrist-watch that utilizes GPS tracking to determine distance and pace. Sensors and components include 3-axis accelerometers, digital compass, optical heart rate monitor, altimeter, ambient light sensor, and vibration motor. The Polar A360 (Polar Electro, Kempele, Finland) is a wrist-worn fitness tracker that has a proprietary optical heart rate module. No other specifications are given. The Garmin Vivosmart HR+ (Garmin Ltd, Canton of Schaffhausen, Switzerland) is smart activity tracker with wrist-based heart rate as well as GPS. Sensors include a barometric altimeter and accelerometer. The Leaf Health Tracker (Bellabeat, San Fransisco, CA): the Nature model is 41 cm long, and weighs 16.5 g. Sensors include a 3-axis accelerometer and vibration motor.

Of the wearable technology devices investigated, four are worn on the wrist: Samsung Gear 2, FitBit Surge, Polar A360, Garmin Vivosmart HR+, and one worn on the waist: Leaf Health Tracker. Immediately prior to testing, participant age, gender, height, weight, and where the device was being worn were programmed into the device. Devices were placed on the wrist in a randomized configuration. The device was synchronized via Bluetooth to an iPad (Apple Inc., Cupertino CA), and the appropriate "activity" mode, if available, was selected. The mean of two

manual step counts using a hand-held tally counter (Horsky, New York, NY) was used as the criterion measurement. All devices use proprietary algorithms to determine what constitutes a step for counting purposes.

Protocol

Data for this study was completed concurrently during a collection period that has been recently published (26). The protocol has been described here for the convenience of the reader. In the week prior to testing, participants provided anthropometric data. Height (cm) was measured with a Health-o-meter wall mounted height rod (Pelstar LLC/Health-o-meter, McCook, IL), mass (kg) and Body Mass Index (BMI) was provided by a hand-and-foot bioelectric impedance analyzer (seca mBCA 514 Medical Body Composition Analyzer, Seca North America, Chino, CA). Age in years was self-reported.

On the first day of testing, participants were fitted with the Samsung Gear 2, FitBit Surge, Polar A360, Garmin Vivosmart HR+ and Leaf Health Tracker. They then proceeded to a long indoor hallway with cones spaced 200 feet apart. Participants sat for 5 minutes and then completed the first 5-minute self-paced free motion walk back and forth between the cones while step count was recorded by the two manual counters. After a 5-minute seated rest period, participants completed the first 5-minute self-paced free motion jog with step count again recorded by two manual counters. Participants then rested in a seated position for 10 minutes. They then performed a second self-paced 5-minute free motion walk and jog in the same manner as the first with step count recorded in the same manner. The two manual counters for all free-motion walks and jogs were positioned near the center of the testing area but were separated so they could not view each other's thumb motion nor hear the "clicking" from the tally counter. This prevented any synchronized counting between the two. The manual counters were instructed not to follow or move with the participants to prevent influencing their walking/jogging speed. The distance traveled for both free motion walks and jogs was measured and the speed in miles per hour was calculated and rounded to the nearest 0.1. The calculation used was (feet traveled/5,280)* 12.

One to 2 days later at approximately the same time of day (±1 hour), the participants returned for treadmill-based walking and jogging. They were fitted with all the devices in the same manner and configuration as on day two. All treadmill activities were performed on a Trackmaster treadmill (Full Vision, Inc. Newton, KS). After a 5-minute seated rest period, they completed the first 5-minute treadmill walk at the speed calculated from the first free motion walk with step count recorded by the two manual counters. Following a 5-minute seated rest period, they completed the first 5-minute treadmill jog at the speed calculated from the first free motion jog with step count again recorded by the two manual counters. Participants rested in a seated position for 10 minutes. They then performed a second 5-minute treadmill walk and jog with step count recorded in the same manner as the first treadmill activities. Speeds for the second treadmill walk and jog were calculated from the second free motion walk and jog. Speeds were replicated on the treadmill in order to normalize the distance a participant traveled in the 5-minute testing intervals for both conditions. The grade for all treadmill testing was set to 0%.

The two manual counters were positioned at opposite sides of the lab in order to prevent any synchronized "clicking."

Statistical Analysis

IBM SPSS (IBM Statistics version 24.0, Armonk, NY) was used for all statistical analysis. Three individual data outliers of ≥ ±3 standard deviations were removed from the analysis. These were participant #7 and #14, FitBit Surge, free motion jog #2: step count was not recorded properly at the end of both said activities. Also, participant #37, Samsung Gear 2, treadmill walk #2: device stopped counting and had to be re-synchronized to reset step counting function for next activity. Inter-rater reliability between the two manual counters (n = 40), test-retest of the five devices (n = 40 except for Fitbit Surge, free motion jog: n = 38 and Samsung Gear 2, treadmill walk: n = 39), and validity testing (n = 40 except for Fitbit Surge, free motion jog: n = 38 and Samsung Gear 2, treadmill walk: n = 39) was calculated for free motion walking, free motion jogging, treadmill walking, and treadmill jogging.

The first and second walks and first and second jogs for both the free motion and treadmill activities were compared to one another for reliability. Inter-rater and test-retest reliability were determined using Intraclass Correlation Coefficient (ICC; Model 3, single rating) with an ICC ≥ 0.70 being acceptable (4). The second walk and second jog for the free motion and treadmill activities were used for determining both the standard error of difference (SEd) between the two manual counters and for validity testing. Validity was determined using 1) the mean of the two manual step counters and 2) the values obtained from the wearable technology devices. Pearson's correlation coefficient (r) was used to determine criterion validity with the p-value set at < 0.05 and the (r) set at ≥ 0.70. Secondly, mean absolute percentage error (MAPE) was calculated by the formula: absolute value of ([mean difference of device – comparison] * 100) / comparison mean. Based on previous studies, an acceptable mean absolute percent score is ≤ 10% in free motion movement and ≤ 5% on a treadmill (30, 31, 32). Lastly, a Bland-Altman analysis was performed to help ascertain if the device had a high or low bias in its measurements. Because there are no current guidelines for what an acceptable limit of agreement value would be for a wearable technology device, our results were reported for potential future meta-analysis. Confidence intervals were set at 95%.

RESULTS

Speed and step count are reported in Table 1.

Table 1. Speed and step count (± standard deviation)

Activity (n = 40)	Speed (mph)	Manual Counters (mean)	Samsung Gear 2	FitBit Surge	Polar A360	Garmin Vivosmart HR+	Leaf Health Tracker
Free Motion Walk 1	2.8	560 ± 44	532 ± 70	535 ± 43	522 ± 60	557 ± 44	561 ± 74
Free Motion Walk 2	2.8	563 ± 42	537 ± 55	536 ± 38	529 ± 54	558 ± 42	572 ± 38
Free Motion Jog 1	4.9	792 ± 43	749 ± 138	792 ± 35	737 ± 68	802 ± 47	793 ± 70
Free Motion Jog 2	4.9	804 ± 37	801 ± 36	802 ± 38*	742 ± 52	805 ± 44	808 ± 40
Treadmill Walk 1	2.8	561 ± 43	536 ± 47	534 ± 51	506 ± 76	557 ± 44	551 ± 93
Treadmill Walk 2	2.8	560 ± 45	522 ± 78**	531 ± 57	506 ± 55	558 ± 44	571 ± 52
Treadmill Jog 1	4.9	787 ± 47	768 ± 51	780 ± 50	736 ± 54	787 ± 50	790 ± 49
Treadmill Jog 2	4.9	796 ± 53	776 ± 65	794 ± 56	735 ± 62	795 ± 53	799 ± 56

Note. * n = 38, ** n = 39

Inter-rater Manual Step Count Reliability and Standard Error of Difference. Manually counted steps by two independent counters were determined to be sufficiently reliable for all four activities (n = 40). The standard error of difference between the two counters was also acceptable. Free motion walk, ICC = 0.99, SEd = 10 steps. Free motion jog, ICC = 0.97, SEd = 9 steps. Treadmill walk, ICC = 0.99, SEd = 10 steps. Treadmill jog, ICC = 0.99, SEd = 12 steps.

Device Reliability and Validity. The Samsung Gear 2 returned significant ICC, p, and (r) values for both jogging activities (Table 2, Figures 1A-1D). However, for both walking activities, while the p-value was significant, the ICC and (r) values were not. Both free motion activities had acceptable mean absolute percent errors (MAPE) of ≤ 10.0% and the treadmill jogging was ≤ 5%. While treadmill walking returned a significant p-value, the ICC and (r) values were not. Also, the MAPE for treadmill walking was unacceptable at > 5%. Bland-Altman plots suggest that it underestimates step count measurements during all activities.

The FitBit Surge returned significant ICC, p, and (r) values for all four activities (Table 3, Figures 2A-2D). Two outliers were removed from the free motion jog analysis. While the mean absolute percent error (MAPE) was acceptable at ≤ 10.0% for both free motion activities and ≤ 5% level for the treadmill jog, the treadmill walk MAPE was unacceptable being slightly higher than 5%. Bland-Altman plots suggest that it underestimates step count measurements for all activities with the walking activities being noticeably higher than the jogging.

Figure 1A.

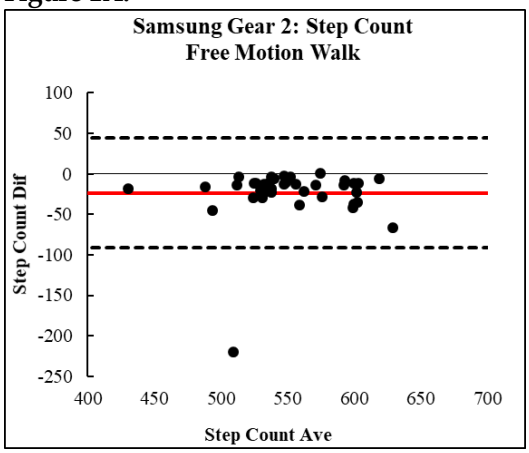


Figure 1B.

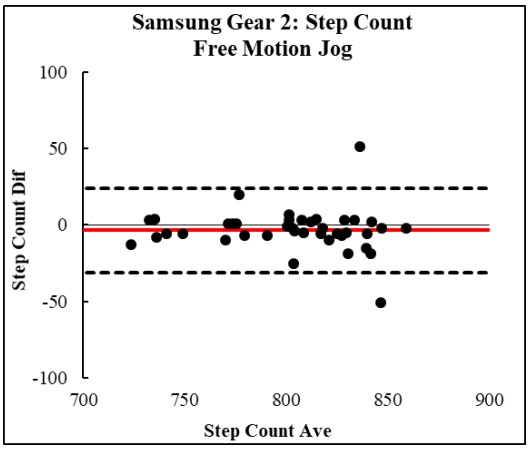


Figure 1C.

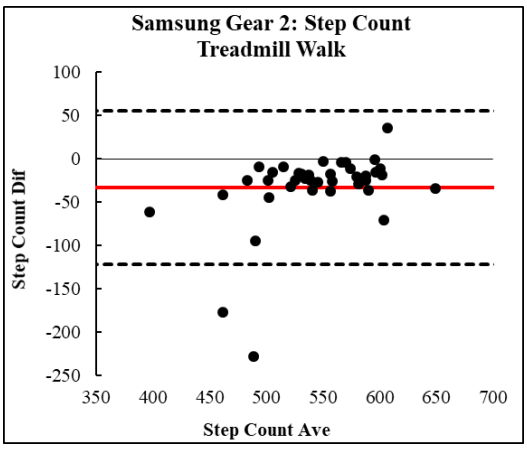
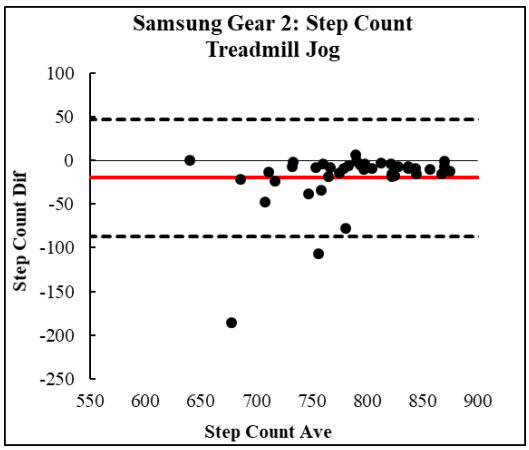


Figure 1D.



Figures 1A. (Free Motion Walk**), 1B. (**Free Motion Jog**), 1C. (**Treadmill Walk**), and 1D. (**Treadmill Jog**).** Samsung Gear 2, Step Count, Bland-Altman plots.

Table 2. Samsung Gear 2. Step Count test-retest and validity.

Reliability (n = 40)		Validity (n = 40)				
	ICC 3,1	r	p	MAPE (%)	Bias (steps)	LoA (steps)
Free Motion Walk	0.57	0.68	0.001	4.09	-24 ± 35	-91 to 44
Free Motion Jog	0.92	0.93	< 0.001	1.08	-3 ± 14	-31 to 24
Treadmill Walk *	0.49	0.54	< 0.001	6.30	-33 ± 44	-122 to 56
Treadmill Jog	0.75	0.85	< 0.001	2.58	-20 ± 34	-87 to 47

Note. ICC = Intraclass Correlation Coefficient, MAPE = Mean Absolute Percent Error, LoA = Limits of Agreement, * n = 39

Table 3. FitBit Surge. Step Count test-retest and validity.

Reliability (n = 40)		Validity (n = 40)				
	ICC 3,1	r	p	MAPE (%)	Bias (steps)	LoA (steps)
Free Motion Walk	0.86	0.83	<0.001	4.84	-27±24	-74 to 19
Free Motion Jog*	0.90	0.92	<0.001	1.42	-1±16	-32 to 29
Treadmill Walk	0.76	0.75	<0.001	5.84	-29±38	-103 to 46
Treadmill Jog	0.84	0.94	<0.001	1.45	-2±9	-39 to 35

Note. * n = 38

Figure 2A.

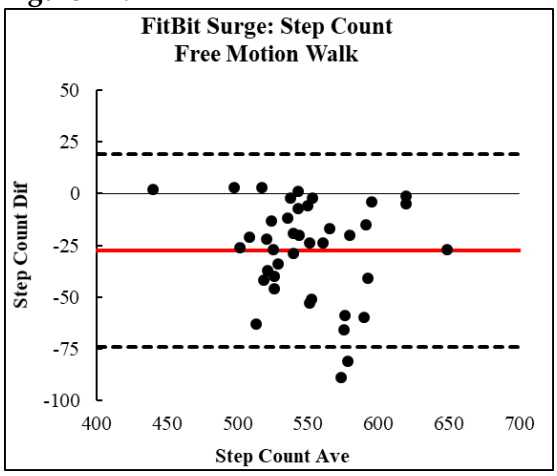


Figure 2B.

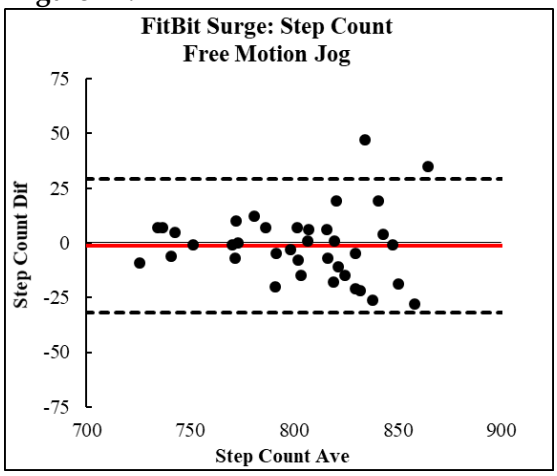


Figure 2C.

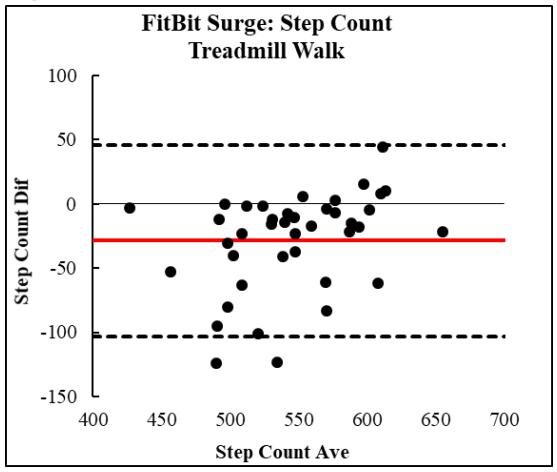
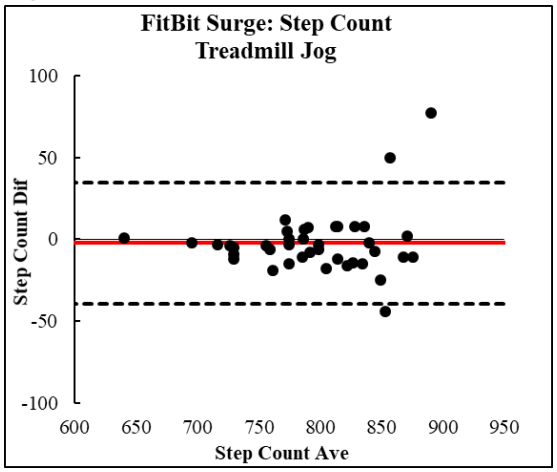


Figure 2D.



Figures 2A. (Free Motion Walk**), 2B. (**Free Motion Jog**), 2C. (**Treadmill Walk**), and 2D. (**Treadmill Jog**).** FitBit Surge, Step Count, Bland-Altman plots.

Figure 3A.

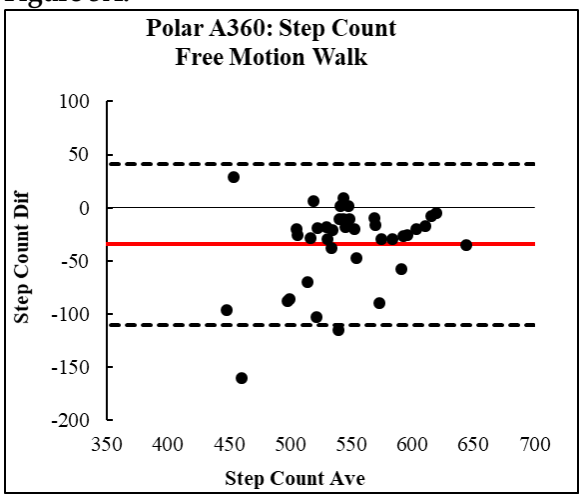


Figure 3B.

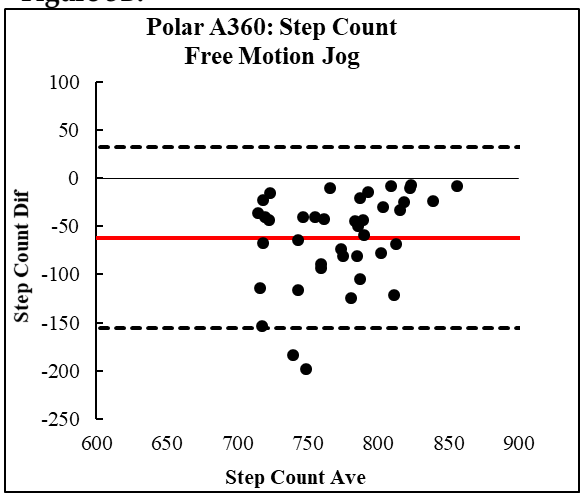


Figure 3C.

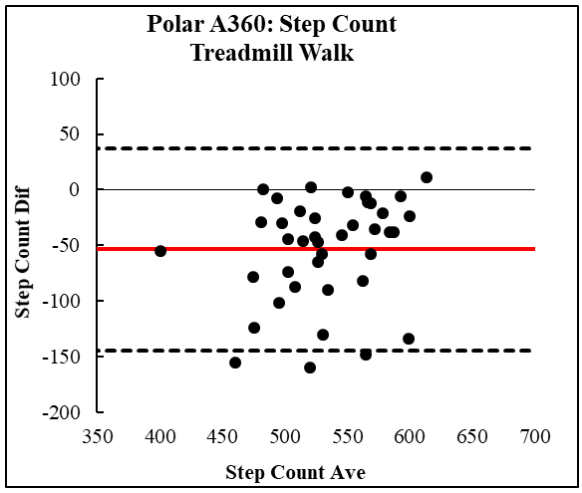
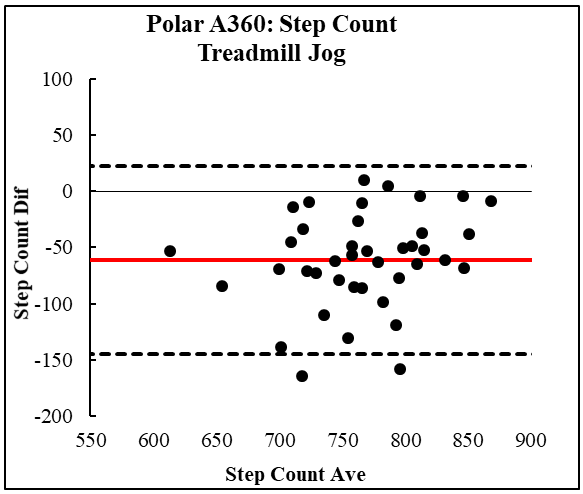


Figure 3D.



Figures 3A. (Free Motion Walk**), 3B. (**Free Motion Jog**), 3C. (**Treadmill Walk**), and 3D. (**Treadmill Jog**).** Polar A360, Step Count, Bland-Altman plots.

Table 4. Polar A360. Step Count test-retest and validity.

Reliability (n = 40)		Validity (n = 40)				
	ICC 3,1	r	p	MAPE (%)	Bias (steps)	LoA (steps)
Free Motion Walk	0.52	0.69	< 0.001	6.58	-34 ± 39	-110 to 41
Free Motion Jog	0.44	0.46	0.003	7.64	-62 ± 48	-156 to 32
Treadmill Walk	0.51	0.59	< 0.001	9.58	-54 ± 46	-145 to 38
Treadmill Jog	0.78	0.74	< 0.001	7.75	-61 ± 42	-145 to 22

The Polar A360 returned significant ICC, p, and (r) values only for treadmill jogging (Table 4, Figures 3A-3D). For both free motion activities and treadmill walking, while the p-value was significant, the ICC and (r) values were not. Both free motion activities had acceptable mean absolute percent errors (MAPE) of ≤ 10.0% while both treadmill activities were unacceptable at

> 5%. Bland-Altman plots suggest that it greatly underestimates step count measurements during all four activities.

The Garmin Vivosmart HR+ returned significant ICC, *p*, and (*r*) values for all four activities (Table 5, Figures 4A-4D). The mean absolute percent error (MAPE) was acceptable for all with ≤ 10.0% for both free motion activities and ≤ 5% for both of those on the treadmill. Bland-Altman plots suggest that it minimally underestimates step count measurements during free motion and treadmill walking, and treadmill jogging. It minimally overestimates step counts when free motion jogging.

Table 5. Garmin Vivosmart HR+. Step Count test-retest and validity.

Reliability (*n* = 40)		Validity (*n* = 40)				
	ICC 3,1	*r*	*p*	MAPE (%)	Bias (steps)	LoA (steps)
Free Motion Walk	0.74	0.81	< 0.001	2.47	-5 ± 26	-56 to 46
Free Motion Jog	0.82	0.87	< 0.001	1.48	1 ± 21	-41 to 44
Treadmill Walk	0.87	0.98	< 0.001	1.36	-2 ± 10	-22 to 18
Treadmill Jog	0.93	0.99	< 0.001	0.56	-1 ± 6	-13 to 11

Figure 4A.

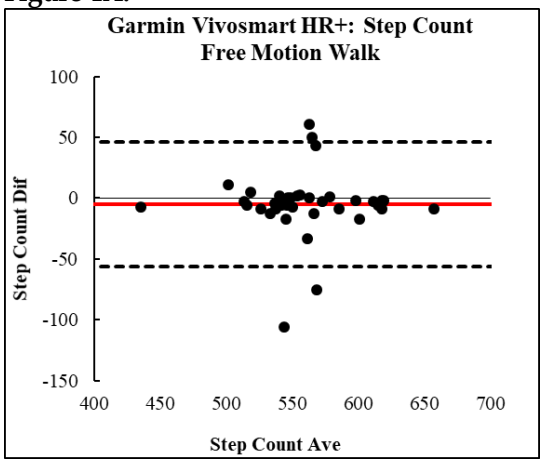


Figure 4B.

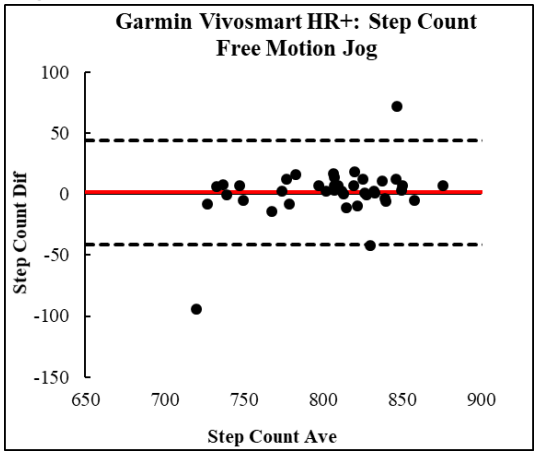


Figure 4C.

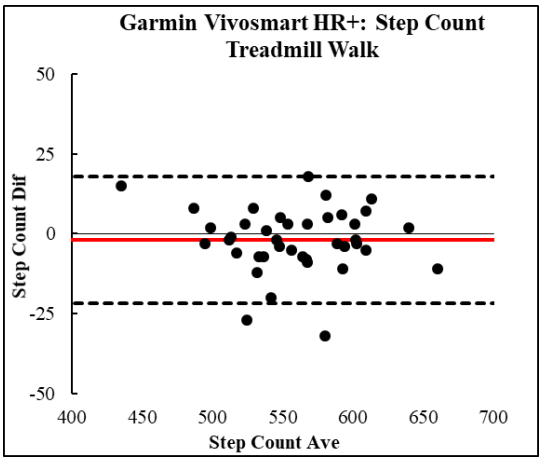
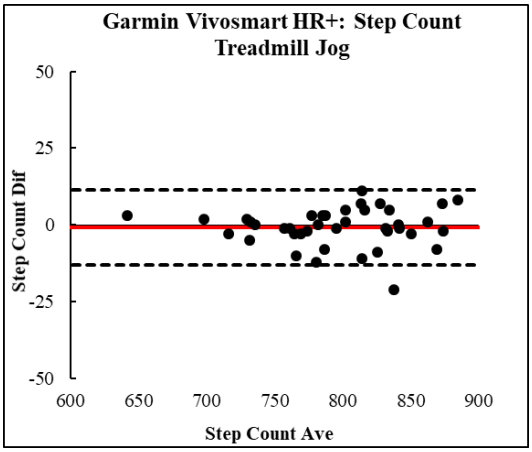


Figure 4D.



Figures 4A. (Free Motion Walk**), 4B. (**Free Motion Jog**), 4C. (**Treadmill Walk**), & 4D. (**Treadmill Jog**).** Garmin Vivosmart HR+, Step Count, Bland-Altman plots.

Figure 5A.

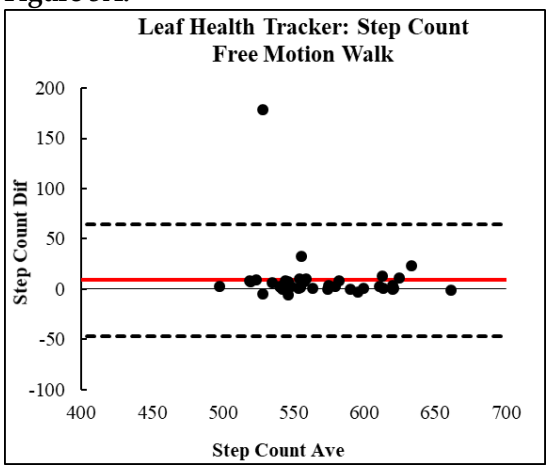


Figure 5B.

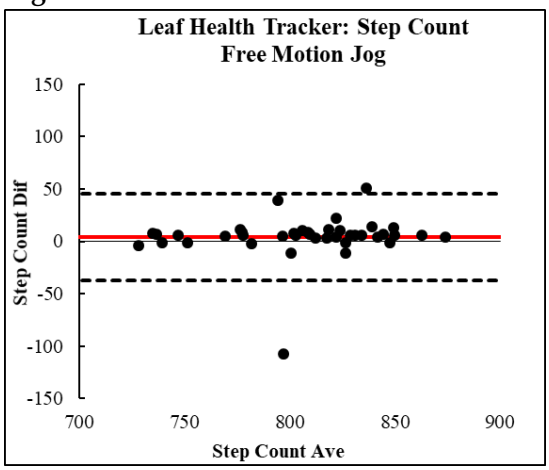


Figure 5C.

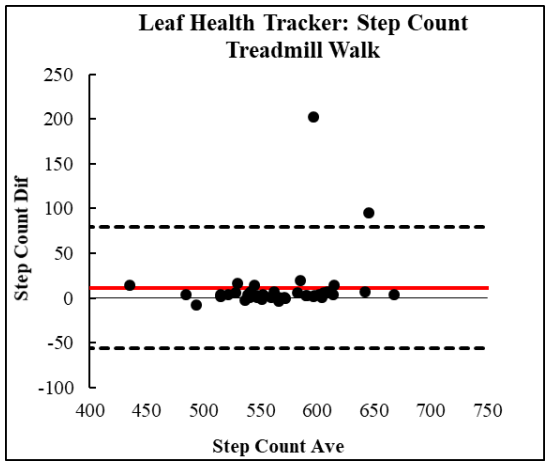
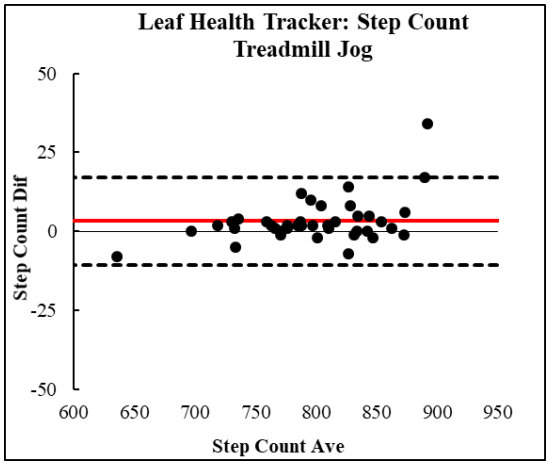


Figure 5D.



Figures 5A. (Free Motion Walk**), 5B. (**Free Motion Jog**), 5C. (**Treadmill Walk**), & 5D. (**Treadmill Jog**).** Leaf Health Tracker, Step Count, Bland-Altman plots.

The Leaf Health Tracker returned significant ICC, *p*, and (*r*) values for all four activities (Table 6, Figures 5A-5D). The mean absolute percent error (MAPE) was acceptable for all with a ≤ 10.0% for both free motion activities and ≤ 5% for both of those on the treadmill. Bland-Altman plots suggest that it minimally overestimates step count measurements for all activities.

Table 6. Leaf Health Tracker. Step Count test-retest and validity.

Reliability (n = 40)		Validity (n = 40)				
	ICC 3,1	r	p	MAPE (%)	Bias (steps)	LoA (steps)
Free Motion Walk	0.72	0.75	< 0.001	1.96	9 ± 28	-47 to 65
Free Motion Jog	0.86	0.85	< 0.001	1.39	4 ± 21	-37 to 46
Treadmill Walk	0.72	0.76	< 0.001	2.30	12 ± 34	-56 to 179
Treadmill Jog	0.93	0.99	< 0.001	0.57	3 ± 7	-11 to 17

DISCUSSION

The current study investigated the accuracy of five wearable technology devices that recorded participant step counts during the activities of walking and jogging. Measurements were taken during walk and jog intervals performed in both a free motion setting and on a treadmill. The criterion measure was the mean of steps recorded by two independent manual counters. The three-fold purpose of this investigation was to determine: 1) step count test-retest reliability of the wearable technology devices for all participants while walking and jogging in both a free motion and treadmill setting, 2) validity of said wearable technology devices for all participants, and 3) evaluate the inter-reliability of two independent manual counters.

In order to be considered valid in the current investigation, a device had to return both an ICC greater than 0.70, and MAPE less than 5% during treadmill exercise, or less than 10% during free motion activity. Of the five devices tested, the Garmin Vivosmart HR+ (Table 5) and Leaf Health Tracker (Table 6.) were observed to be reliable and valid across all tested situations. The FitBit Surge (Table 3) was valid and reliable across all conditions with the exception of treadmill walking. The Samsung Gear 2 (Table 2) was observed to be reliable and valid for the jogging conditions (both treadmill and free motion), but neither of the walking conditions. The Polar A360 (Table 4) was found to be reliable for one condition (treadmill jog), but not valid for any condition. While it is possible that location placement could have affected validity measurements for certain devices, we believe this to not be likely as devices were attached to participants in a randomized order. Some step counting devices have been shown to be less accurate at slower speeds with a necessary threshold at a pace above 3.5 mph (25). A possible explanation for the devices that did not meet the validity threshold in the current study, could be that participants were executing activity at too slow of a pace to return accurate step count. It should be noted that conditions deemed not valid involved primarily walking – FitBit Surge treadmill walking, Samsung Gear 2 treadmill and free motion walking, and Polar A360 treadmill and free motion walking. Researchers and consumers who are relying on the above devices for accurate step count during walking activities should take these findings into consideration.

Wearable technology devices have been tested for step count accuracy in laboratories (1, 24, 25), inside on a track or hallway (16, 30), and on outside paved roads (19). To our knowledge, this is the first investigation to evaluate a wearable technology device for step count measures when walking and jogging in both a free motion setting and on a treadmill.

The Garmin Vivosmart HR+ has been evaluated four previous times that we are aware of with three being laboratory/treadmill based and one using a self-selected speed in an indoor hallway and on an outdoor path (17, 20, 33, 40). For the self-selected speed protocol when walking indoors and outdoors, it was shown to have a low mean absolute percent error for both (< 3%). This was comparable to our study (≤ 2.47%). While this study had consistently high (*r*) values for all outdoor free motion walking (0.94 - 0.97), our study was lower for the same activity (0.74) (20). Laboratory studies found healthy participants running at two different speeds on a treadmill had mean absolute percent errors of < 2% for both activities (39), and when individually evaluated during one mile walks and one mile runs on a treadmill, the Garmin

Vivosmart HR+ was not valid when walking at slower speeds but was valid when running at speeds > 4.5 mph (33). The Garmin Vivosmart HR+ exhibited increasing mean absolute percent errors as the walking speed increased on a treadmill (3.2 km•hr-1 = 1% to 6.4 km•hr-1 = 9%). Our results showed a mean absolute percent error of 1.36% for treadmill walking (Table 5).

The FitBit Surge has been evaluated in four studies utilizing both a treadmill and in a free motion setting (5, 22, 28, 41) When compared to an Apple Watch and the Microsoft Band, the FitBit Surge showed the most discrepancy when related to a criterion measurement for both treadmill walking and treadmill jogging at different speeds (5). During a 5-day free motion/living study, numerous devices, including the FitBit Surge, were shown to have an ICC of 0.89. However, no criterion measure was reported (41). In a study where participants walked 1,000 steps, the FitBit Surge underestimated step count for all age groups tested (22). This was in line with our study where the FitBit Surge appeared to underestimated step count for all four of our testing settings. The FitBit Surge was shown to be valid while walking during trail hiking but that the accuracy worsened as the activity become more intense (28). Our results show that with one slightly high exception in the mean absolute percent error (5.84%), the FitBit Surge is both reliable and valid when walking or jogging (Table 3).

The Samsung Gear 2 was found to be evaluated in three studies (14, 22, 28). In a study where participants walked 200, 500, and 1,000 steps, the Samsung Gear 2 overestimated steps in every trial (14). In a different study where participants only walked 1,000 steps, it underestimated steps for a 40-64 year old age group (22). Our study showed that the Samsung Gear 2 underestimated step count for all four situations tested. Step count measured during a trail hiking and running study saw inconsistent results as the hiking ICC and running mean absolute percent error were accurate but hiking mean absolute percent error and running ICC were not (28). Our study reported a large underestimation of step count measures in contrast to this study which reported the Samsung Gear 2 overestimated step count in all cases (Table 2).

The Polar A360 has only two known published studies (6, 28). During a self-selected walking and running protocol on a treadmill at 1% grade, the Polar A360 underestimated the treadmill walking step count but had an acceptable mean absolute percent error (< 5%). However, during treadmill running, step count underestimation increased with the mean absolute percent error increasing to well above acceptable levels (> 10%). Our results indicated a large underestimation of step count for all four conditions with both the treadmill walk and jog having mean absolute percent errors above 5% (Table 4). In contrast, trail running analysis revealed an overestimation of step count (28).

Even though it has been mentioned in the literature (2, 13, 32), there is only one known study that has evaluated the Leaf Health Tracker (28). During a trail running setting, it was shown to have an (r) = 0.95 with a small underestimation of step count. Our results were similar in that the (r) values were acceptable for all activities. In contrast though, we saw an overestimation of step count for every condition (Table 6).

We are aware that there is abundant literature on the validation of wearable technology but very little on test-retest reliability. Systematic reviews have identified a pattern whereas researchers are simply validating wearable technology devices without determining reliability (7, 15). It can be speculated that this can be attributed to a sense of urgency by researchers to get information out to the public quickly. Because the field of wearable technology is rapidly evolving and expanding, by the time a product is tested and the results released, that product may already have been upgraded or replaced. Also, because recruiting and retaining participants for reliability purposes is more difficult and time consuming, investigators may not have the ability to do so. Either way, this incomplete analysis can be deceptive. Reliability, being a component of validity, means that without test-retest analysis, a wearable technology device cannot truly be considered as valid for accuracy purposes. We purposefully designed our study to account for this.

One of the purposes for this study was to determine if the mean of two independent manual counters could be a practical criterion measure when evaluating device step count values. The flexibility and mobility of two manual step counters is practical in most situations and would give instantaneous results as opposed to evaluating the data at a later point in time. Additionally, manual counters would not require an investment in equipment to record and watch a video later. This would both potentially save time and keep costs low. Finally, it can be argued that counting steps for a live participant would retain a counter's attention more than having to sit in front of a monitor and watch a video. Video watching, while simple, can be boring and repetitious. These factors may result in the watchers miscounting due to being inattentive and therefore not reporting the exact same step counts as required. Manual step counts by a single counter (1, 17, 24, 25) and two counters (16, 28) have already been used as a criterion measure. For the two previous studies that used dual manual counters, the inter-rater reliability was > 0.99 for all protocols analyzed. We can add to the literature using two counters as our lowest inter-rater reliability value was 0.97 (free motion jogging) with all others being > 0.99.

In summary, the purposes of this investigation were to determine step count reliability and validity of wearable technology devices in free motion and treadmill settings and to evaluate the inter-reliability of two manual counters as a basis for use as a criterion measure. We presented strong evidence that two independent manual counters have a high inter-reliability correlation. Two counters could reasonably be used as a sound methodology for step count protocols as the criterion measure. We also found that overall, except for the Samsung Gear 2 and the Polar A360, that the wearable technology devices tested were acceptable for use in daily step counts.

This study only evaluated step counts measured by the devices. While this is important for obtaining and maintaining a healthy lifestyle, it is not the only factor that needs to be addressed for these purposes. Future research should also examine the consistency and accuracy of wearable technology to estimate energy expenditure, or calorie consumption, as either a separate factor or in conjunction with step counts. For example, a device that over estimates both step count and estimated energy expenditure can create an unfortunate situation where the

wearer will believe they are performing the recommended amount of daily physical activity and burning more calories than they really are. Users may not see the anticipated weight loss or physiological improvements over time that should correlate with the devices recorded values. This can cause frustration and demoralize them from continuing, causing them to stop due to no fault of their own.

As the use of wearable technology devices becomes more prevalent for controlling obesity rates and promoting healthy lifestyles, their accuracy and consistency must be evaluated in as many real-life settings and populations as possible. While we only evaluated four activity situations, the average person does far more than that in their daily life. Constraining our investigation to only these activities could be considered a limitation of this study. Motions such as using stairs to transverse floors in a building, bending and reaching motions, riding stationary and standard cycles, and the use of swimming pools or elliptical machines in a gym all present new movement patterns that will also require evaluation and incorporation into the measurement of daily activity levels. Our participants were mostly young, fit college students. Future research should examine if a physically unfit or older population will have different results than that produced by this study. Obese and elderly persons may have different motion mechanics that could result in increased inaccuracy. Being that these populations could potentially be more reliant on these devices to achieve healthier levels of physical fitness, it is important that they give as precise of measurements that they can for them.

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. Balaam M., Hansen LK, D'Ignazio C, Simpsom E, Almeida T, Kuznetsov S, Catt M, Søndergaard MLJ. Hacking Women's Health. Paper presented at the Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA, 2017.

3. Bassett Jr DR., Toth LP, LaMunion SR, Crouter SE. Step Counting: A Review of Measurement Considerations and Health-Related Applications. Sport Med 47(7): 1303-1315, 2017.

4. Baumgartner TA., Jackson AS, Mahar MT, Rowe DA. Measurements for evaluation in physical education and exercise science (8th ed.). New York, NY: McGraw Hill, 2007.

5. Binsch O, Wabeke T, Valk PJ. Comparison of three different physiological wristband sensor systems and their applicability for resilience- and work load monitoring. Paper presented at the IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks, San Francisco CA, 2016.

6. Bunn JA, Jones C, Oliviera A, Webster MJ. Assessment of step accuracy using the Consumer Technology Association standard. J Sport Sci: 1-5, 2018.

7. Bunn JA, Navalta JW, Fountaine CJ, Reece JD. Current State of Commercial Wearable Technology in Physical Activity Monitoring 2015-2017. Inter J Exerc Sci 11(7): 503-515, 2018.

8. CDC. (08/13/2018). Overweight & Obesity. Adult Obesity Facts. Retrieved from https: //www.cdc.gov/obesity /data/adult.html

9. CfHS, N. (2010). Health, United States, 2009: With Special Feature on Medical Technology. Hyattsville MD: Med Tech.

10. 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.

11. Choi BC, Pak AW, Choi JC, Choi EC. Daily step goal of 10,000 steps: a literature review. Clin Invest Med 30(3): E146-151, 2007.

12. Consumer Technology Association. Physical Activity Monitoring for Fitness Wearables: Step Counting. Arlington VA: Consumer Technology Association, 2016

13. Eatough E, Shockley K, Yu P. A Review of Ambulatory Health Data Collection Methods for Employee Experience Sampling Research. J Appl Psych 65(2): 322-354, 2016.

14. El-Amrawy F, Nounou MI. Are Currently Available Wearable Devices for Activity Tracking and Heart Rate Monitoring Accurate, Precise, and Medically Beneficial? J Healthc Inform Res 21(4): 315-320, 2015.

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

16. Floegel TA., Florez-Pregonero A, Hekler EB, Buman MP. Validation of Consumer-Based Hip and Wrist Activity Monitors in Older Adults with Varied Ambulatory Abilities. J Gerontol A Biol Sci Med Sci 72(2): 229-236, 2017.

17. Fokkema T, Kooiman TJ, Krijnen WP, Van Der Schans CP, Groot, M. Reliability and Validity of Ten Consumer Activity Trackers Depend on Walking Speed. Med. Sci Sports Exerc 49(4): 793-800, 2017.

18. Hatano Y. Use of the pedometer for promoting daily walking exercise. Int Council Health, Phys Educ, and Rec 29: 4-8, 1993.

19. 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.

20. Lamont RM, Daniel HL, Payne CL, Brauer SG. Accuracy of wearable physical activity trackers in people with Parkinson's disease. Gait Posture 63: 104-108, 2018.

21. Liguori G, Dweyer GB, Fitts TC. ACSM:. Resources for the Health Fitness Specialist (1st ed.). Philadelphia PA: Lippincott, Williams, and Wilkins, 2014.

22. Modave F, Guo Y, Bian J, Gurka MJ, Parish A., Smith MD, Lee AM, Buford TW. Mobile Device Accuracy for Step Counting Across Age Groups. JMIR Mhealth and Uhealth 5(6): e88, 2017. doi:10.2196/mhealth.7870

23. Montes J, Stone TM., Manning JW, McCune D, Tacad DK, Young JC, Debeliso M, Navalta JW. Using Hexoskin Wearable Technology to Obtain Body Metrics During Trail Hiking. Int J Exerc Sci 8(4): 425-430, 2015.

24. Montes J, Young, JC, Tandy R., Navalta JW. Fitbit Flex: Energy Expenditure and Step Count Evaluation. J Exerc Physiol online, 20(5): 152-159, 2017.

25. Montes J, Young JC, Tandy R., Navalta JW. Reliability and Validation of the Hexoskin Wearable Bio-Collection Device During Walking Conditions. Int J Exerc Sci 11(7): 808-816, 2018.

26. Montes J, Navalta JW. Reliability of the Polar T31 Uncoded Heart Rate Monitor in Free Motion and Treadmill Activities. Int J Exerc Sci 12(4): 69-76, 2019.

27. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK., Blaha MJ, Cushman M. Stroke Statistics, Heart disease and stroke statistics--2015 update: a report from the American Heart Association. Circulation 131(4): e29-322, 2015.

28. Navalta JW, Montes J, Bodell NG, Aguilar CD, Lujan A, Guzman G, Kam BK. Wearable Device Validity in Determining Step Count During Hiking and Trail Running. J Measure Phys Behavior 1: 86-93, 2018.

29. Navalta JW, Stone WJ, Lyons TS. Ethical Issues Relating to Scientific Discovery in Exercise Science. Int J Exerc Sci 12(1): 1-8, 2019.

30. 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.

31. Schneider PL, Crouter S, Bassett DR. Pedometer measures of free-living physical activity: comparison of 13 models. Med. Sci Sports Exerc, 36(2): 331-335, 2004.

32. Silina Y, Haddadi H. New directions in jewelry. Paper presented at the Proceedings of the 2015 ACM International Symposium on Wearable Computers - ISWC 2015.

33. Smit MA, Powers M. Does the Garmin Vivosmart HR+ accurately measure steps and energy expenditure? Int J Exerc Sci Conference Abstracts 11(4): 2016

34. Statista. (2018a, 06/01/2018). Forecast wearables unit shipments worldwide from 2014 to 2022 (in millions). Retrieved from https://www.statista.com/statistics/437871/wearables-worldwide-shipments/

35. Statista. (2018b, 09/01/2017). Wearable device sales revenue worldwide from 2016 to 2022 (in billion U.S. dollars). Retrieved from https://www.statista.com/statistics/610447/wearable-device-revenue-worldwide/

36. Thompson WR. Worldwide Survey of Fitness Trends for 2016 10th Anniversary Edition. ACSM Health & Fitness Journal 19(6): 9-18, 2015.

37. Thompson WR. Worldwide Survey of Fitness Trends for 2017. ACSM Health & Fitness Journal 20(6): 8-17, 2016

38. Tigbe WW, Granat MH, Sattar N, Lean MEJ. Time spent in sedentary posture is associated with waist circumference and cardiovascular risk. Int J Obes (Lond) 41(5): 689-696, 2017

39. Tudor-Locke C, Johnson WD, Katzmarzyk PT. Accelerometer-determined steps per day in US adults. Med. Sci Sports Exerc 41(7): 1384-1391, 2009.

40. Wahl Y, Duking P, Droszez A, Wahl P, Mester J. Criterion-Validity of Commercially Available Physical Activity Tracker to Estimate Step Count, Covered Distance and Energy Expenditure during Sports Conditions. Front Physiol 8: 725, 2017.

41. Wen D, Zhang X, Liu X, Lei J. Evaluating the Consistency of Current Mainstream Wearable Devices in Health Monitoring: A Comparison Under Free-Living Conditions. J Med Internet Res 19(3): e68, 2017.