International Journal of EXERCISE SCIENCE

Review

Measuring Physical Activity Using Triaxial Wrist Worn Polar Activity Trackers: A Systematic Review

ANDRÉ HENRIKSEN†1, JONAS JOHANSSON‡1, GUNNAR HARTVIGSEN‡2, SAMELINE GRIMSGAARD‡1, and LAILA HOPSTOCK‡1

[1]Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, NORWAY; [2]Department of Computer Science, UiT The Arctic University of Norway, Tromsø, NORWAY

†Denotes graduate student author, ‡Denotes professional author

ABSTRACT

International Journal of Exercise Science 13(4): 438-454, 2020. Collecting objective physical activity data from research participants are increasingly done using consumer-based activity trackers. Several validation studies of Polar devices are conducted to date, but no systematic review of the current level of accuracy for these devices exist. The aim of this study is therefore to investigate the accuracy of current wrist-worn Polar devices that equips a triaxial accelerometer to measure physical activity. We conducted a systematic review by searching six databases for validation studies on modern Polar activity trackers. Studies were grouped and examined by tested outcome, i.e. energy expenditure, physical activity intensity, and steps. We summarized and reported relevant metrics from each study. The initial search resulted in 157 studies, out of which fourteen studies were included in the final review. Energy expenditure was reviewed in seven studies, physical activity intensity was reviewed in four studies, and steps was reviewed in 11 studies. There is a large difference in study protocols with conflicting results between the identified studies. However, for energy expenditure there is some indication that Polar devices perform better in free-living, compared to lab-based studies. In addition, step counting seems to have less average error compared to energy expenditure and physical activity intensity. There is large heterogeneity between the identified studies, both in terms of study protocols and results, and the accuracy of Polar devices remains unclear. More studies are needed for more recently developed devices, and future studies should take care to follow guidelines for assessment of wearable sensors designed for physical activity monitoring.

KEY WORDS: Motor activity, fitness trackers, sports, exercise, energy expenditure, watch, physical activity intensity, steps

INTRODUCTION

Accelerometers have been used to collect objective measurements on physical activity (PA) in research for several years. Although traditionally collected using expensive and sometimes intrusive research grade instruments, studies are increasingly taking advantage of the growing list of available consumer-based activity trackers. These may be less accurate, but are generally cheaper, provides less burden on participants, and can be worn for a longer time-period.

A recent systematic review of wearable activity trackers used in research, by Shin et al. (36), identified 463 articles published between 2013 and 2017 that included a consumer-based activity tracker in their protocol. In addition, two recent systematic reviews with meta-analysis found that including an activity tracker in PA interventions is likely to increase daily step count, energy expenditure (EE), and minutes spent in moderate and vigorous PA (4), and a modest short-term increase in PA may be achieved when using a wearable activity tracker as part of an RCT protocol (41). Newer models appear on the consumer market every year, boasting improved accuracy, additional sensors, and better user experience (18). The number of sold activity trackers are expected to grow from approximately 125 million units in 2018 to 190 million units in 2022 (43). We can therefore expect an increase in number of wearable enhanced research studies in upcoming years.

A few previous systematic reviews have assessed the validity of activity trackers. Evenson et al. (11) reviewed Fitbit and Jawbone activity trackers in 2015, Feehan et al. (13) published a systematic review of Fitbit activity trackers in 2018, Straiton et al. (39) did a systematic review in 2018 on validation studies conducted on participants aged 65 or above, and Bunn et al. (7) conducted a systematic review of validation studies in 2018 on Fitbit, Garmin, Apple, Misfit, Samsung, TomTom, and Lumo devices. Step validity was often high, but not always, and EE and PA intensity were often under- or overestimated depending on study setting.

EE, PA, and steps can be measured using different techniques. The gold-standard for measuring EE in free-living is the doubly labeled water (DLW) method. However, this is an expensive technique, and indirect calorimetry (IC) is currently the most commonly used method for measuring EE in both lab settings and free-living. IC converts measurement of oxygen consumption into EE and is the best alternative to DLW (20). It is also possible to use accelerometers to estimate EE, where activity counts are used to calculate EE using existing and accepted cut-points. PA intensity is estimated using accelerometers, also using existing activity cut-points and converted into minutes of e.g. light-, moderate-, and vigorous PA. Freedson et al. (15) for uniaxial accelerometers and Sasaki et al. (33) for triaxial accelerometers cut-points are most often used when estimating EE or PA in adults. The current gold-standard for measuring steps in lab settings is manual counting. Under free-living conditions pedometers and accelerometers are used to estimate steps (40). In contrast, how consumer-based wearable activity trackers estimates these metrics is mostly unknown and considered company secrets.

Polar (Polar Electro, Finland) was founded in 1977 and has been a leading brand for consumer-based activity trackers and heart rate data collection via a chest worn strap with an electrocardiography monitor. Today Polar offer a range of activity tracker equipment utilizing different sensors, including accelerometers, gyroscopes, electrocardiography (heart rate), photoplethysmography (pulse), and global positioning systems.

Although Bunn et al. (7) and Straiton et al. (39) included some studies that reviewed a Polar device, we could not find any reviews that aimed to specifically include Polar devices in their search. However, in a previous study Polar was identified as one of the top five brands used in research (18). There is a need to systematically review validation studies conducted on modern

Polar activity trackers. The purpose of this review was therefore to investigate the accuracy of current wrist-worn Polar devices that equips a triaxial accelerometer to measure EE, minutes in different PA intensity levels, and steps.

METHODS

We performed a literature search on June 23, 2019, using MEDLINE (Ovid), EMBASE (Ovid), CINAHL, Web of Science, and Scopus. Because the word "Polar" will result in hits in multiple disciplines, we made keyword search variations for each database that specified all Polar wrist-worn triaxial accelerometer based models: "Polar and (A300 or A360 or A370 or M200 or M400 or M430 or M600 or Polar Loop or Loop Crystal or V800 or Vantage)". We combined this with an additional search to limit results to validation studies: "Validate or validation or validity or accuracy or comparison". In addition to database searches, we conducted a reference search among included studies. We also initially included a study by the authors of the current study, which at the time only was available as a preprint (17) (now published).

In order for a study to be included in the final review, it had to 1) examine EE, time spent at different PA intensity levels, or steps, 2) include an objective criterion measure, 3) assess a smart phone compatible wrist-worn triaxial Polar activity tracker, and 4) include an analysis of effect size or error prediction. Non-English articles, review articles, and abstracts were excluded. We also excluded studies that only examined heart rate, because this metric is not based on accelerometer data, and only some models have the required sensor.

After merging results from each database and studies identified through other sources, we removed duplicates. For remaining studies, titles and abstracts were reviewed to determine relevance in accordance with the selection criteria. A full-text review was conducted on remaining studies to assess eligibility. Two investigators independently performed the search and evaluated the retrieved studies for inclusion. Differences between investigators were resolved through discussion. In the case of disagreement, a third author was brought in to support the decision-making.

We grouped studies into categories based on which variable they tested. Three categories were defined: 1) EE, 2) PA intensity levels, and 3) steps. For each study, results were divided into studies performed in lab settings and studies performed under free-living conditions. Furthermore, for each study we extracted effect sizes, prioritizing Pearson's correlation coefficient and intraclass correlation coefficient (ICC), as well as mean absolute percentage error (MAPE) to predict accuracy between device and criterion. We consider a MAPE above 3% as high for studies conducted in lab-settings (35), and MAPE above 10% as high for studies conducted under free-living conditions (34).

Included studies were assessed for risk of bias using a modified criterion validity subscale (42) from the Consensus-Based Standards for the Selection of Health Status Measurement Instruments (COSMIN) checklist (25). This subscale evaluates studies for missing data report, missing data handling, acceptable sample size, acceptable criterion used, methodological design flaws, and acceptable accuracy assessment. Studies were deemed excellent, good, fair, or poor, in each area.

This research was carried out fully in accordance to the ethical standards of the International Journal of Exercise Science (26).

RESULTS

Study selection: The initial search resulted in 157 studies, where 70 were duplicates and removed. After screening title and abstract of the remaining 87 studies, we removed an additional 69 studies. One was a review, two were abstracts, 20 did not investigate validity of physical activity, 43 did not investigate Polar activity trackers, one investigated an older Polar model outside the scope of this paper, and two studies did not use an objective criterion. We assessed the remaining 18 studies for eligibility by performing a full-text review, after which an additional four studies were removed. One study did not investigate Polar activity trackers and three studies investigated the validity of Polar models outside the scope of this paper (i.e. not wrist-worn triaxial). Fourteen studies met the final inclusion criteria and were included in the review. The PRISMA breakdown is given in Figure 1.

Study characteristics: The most frequently used model was the Polar Loop (1, 5, 14, 37, 38, 45, 46), used in seven studies. The Polar V800 (19, 30) and Polar A360 (3, 6) were used in two studies each, and the Polar A300 (2), Polar M600 (9), and Polar M430 (17) were used in one study. The sample size ranged from nine to 95 participants. In all but two studies, participants were healthy volunteers. One study (2) was based on data from participants with chronic obstructive pulmonary disease, and one study (38) included participants with lower-limb prosthesis only. Table 1 gives a summary of study characteristics, including Polar model tested, criterion measurement used, study setting (lab or free-living), and outcome of interest. Seven, four, and 11 studies investigated validity of EE (2, 3, 5, 17, 19, 30, 45), PA (2, 9, 17, 19), and steps (1, 2, 6, 9, 14, 17, 19, 37, 38, 45, 46), respectively.

In total there were 456 participants in the 14 studies, with sample size ranging between 18-95 participants, age ranging between 18-90 years, and height and weight (among studies who reported this) ranging from 150-196 centimeters and 42-125 kilograms, respectively. Table 2 summarizes participant characteristics, including number of participants, and mean, standard deviation, and range for age, body mass index (BMI), weight, and height, when available.

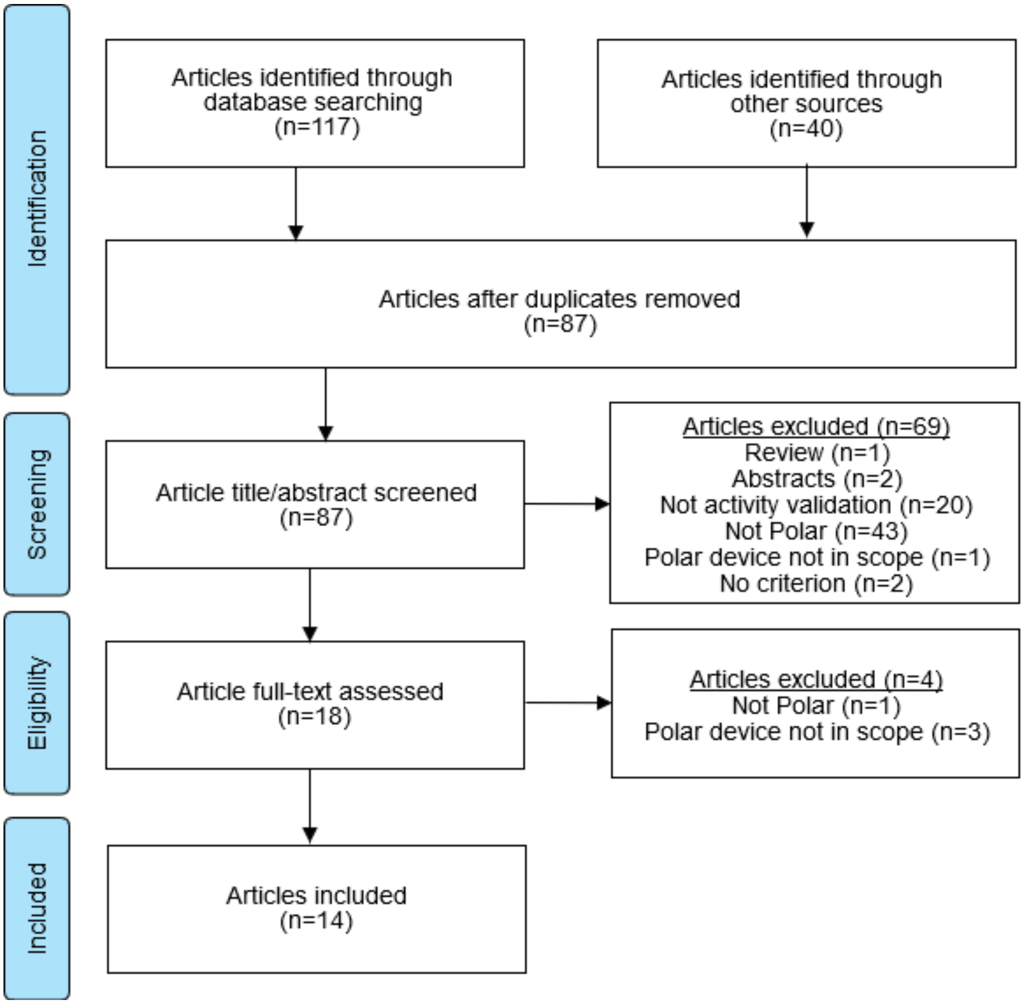
Identification
Screening
Eligibility
Included
Articles identified through database searching (n=117)
Articles identified through other sources (n=40)
Articles after duplicates removed (n=87)
Article title/abstract screened (n=87)
Articles excluded (n=69)
Review (n=1)
Abstracts (n=2)
Not activity validation (n=20)
Not Polar (n=43)
Polar device not in scope (n=1)
No criterion (n=2)
Article full-text assessed (n=18)
Articles excluded (n=4)
Not Polar (n=1)
Polar device not in scope (n=3)
Articles included (n=14)

Figure 1. PRISMA flowchart

Risk of bias: The majority of studies investigating EE, PA intensity levels, and steps against the criterion, showed excellent reporting and handling of missing data (71-100%). In contrast, 50-57% of studies investigating these parameters involved less than 30 participants and were thus deemed poor in the sample size category (13). Only one of the included studies reported their power calculation. For evaluation of steps and PA, most studies (75-82%) scored excellent in using an acceptable criterion, while 43% of the studies investigating EE scored similarly. The remaining 57% of EE studies was deemed good or poor in terms of using an acceptable criterion. Concerning methodology and design flaws, 82% of studies investigating steps scored good-to-excellent, while the same estimation for EE outcome was true for roughly half of the studies. In contrast, 75% of studies investigating PA intensity levels scored below excellent in terms of design flaws. Lastly, almost all studies (> 80%) received good or excellent scores for acceptable means of accuracy estimation for EE, PA, and steps. A summary table for the bias risk assessment is given in Table 3.

Tables 4-6 gives summaries of statistics used in studies investigating EE, PA intensity, and steps, respectively. Each table present available effect sizes, error estimates, and 95% confidence intervals, for each study, with one row for each sub-validation performed. For studies using more than one criterion, we only list results for the criterion currently consider the gold standard, as suggested by Düking et al. (10).

Table 1. Study characteristics for all studies

Study	Model	Criterion	Setting	Outcomes
Hernandez-Vicente et al. (19)	V800	ActiGraph ActiTrainer	Free-living, 7 days	S, EE, PA
Roos et al. (30)	V800	Indirect calorimetry, Moxus Modular Metabolic System	Lab, multiple percentage of VO_2 peak	EE
An et al. (1)	Loop	Manual count, New Lifestyle pedometer	Lab, walking/running multiple speeds/Free-living, 1 day	S
Brook et al. (5)	Loop	Bodymedia SenseWear armband mini	Free-living, 1 day	EE
Fokkema et al. (14)	Loop	Manual count	Lab, walking/running multiple speeds	S
Simunek et al. (37)	Loop	Yamax Digiwalker SW-701, ActiGraph GT3X+	Free-living, 7 days	S
Wahl et al. (45)	Loop	Manual counting and Optogait (S), Metamax 3B (IC)	Lab, walking/running multiple speeds	S, EE
Wang et al. (46)	Loop	Manual count	Lab, multiple intensities and movement types	S
Smith et al. (38)	Loop	Manual count	Lab, 140-meter walk, patients with LLP	S
Boudreaux et al. (3)	A360	IC, TrueOne 2500.	Lab, ergometer cycle exercise test	EE
Bunn et al. (6)	A360	Manual count	Lab, running and walking	S
Boeselt et al. (2)	A300	Bodymedia SenseWear	Free-living, 3 days, patients with COPD	S, EE, PA
Degroote et al. (9)	M600	ActiGraph GT3X+	Free-living, 1 day	S, PA
Henriksen et al. (17)	M430	ActiGraph GT3X-BT, Actiheart 4	Free-living, 1 day	S, EE, PA

Note: S: steps, EE: Energy Expenditure, PA: physical activity intensity, LLP: lower limb prosthesis, COPD: chronic obstructive pulmonary disease, IC: indirect calorimetry

Table 2. Participant characteristics for all studies.

Study	Sample size (M/F). N=456	Age, years (SD)	Age range	Weight, kg (SD)	Weight range	Height, cm (SD)	Height range	BMI, kg/m² (SD)
Hernandez-Vicente et al. (19)	18 (9/9)	21.0 (1.2)	19-23	64.1 (7.9)	53-80	170 (6.6)	158-182	22.3 (1.8)
Roos et al. (30)	20 (12/6)	23.9 (1.9)		66.9 (10.0)		174 (0.1)		
An et al. (1)	35 (17/18)	31.0 (11.8)	19-65	71.7 (13.4)		173 (7.3)		23.8 (3.1)
Brook et al. (5)	95 (34/61)	28.5 (9.9)	19-60	78.6 (14.2)	43-109	174 (9.6)	154-196	25.7 (3.4)
Fokkema et al. (14)	31 (15/16)	32.0 (12.0)						22.6 (2.4)
Simunek et al. (37)	20 (14/6)	24.0 (6.3)						24.3 (4.0)
Wahl et al. (45)	20 (10/10)	25.2 (2.5)		70.7 (13.7)		175 (10.0)		
Wang et al. (46)	9 (5/4)	22.0 (1.0)		58.0 (10.0)		169 (10.0)		
Smith et al. (38)	32 (21/11)	49.7 (14.0)		87.8 (21.1)		171.1 (11.6)		28.1 (5.3)
Boudreaux et al. (3)	50 (22/28)	22.4 (2.9)		76.9 (17.7)		170 (10.6)		26.4 (4.4)
Bunn et al. (6)	20 (10/10)	26.6 (11.5)	18-65	75.8 (19.3)		163 (34.0)		26.4 (6.3)
Boeselt et al. (2)	20 (17/3)	66.4 (7.4)	40-90					28.9 (5.4)
Degroote et al. (9)	36 (16/18)	39.4 (17.8)	20-65	68.4 (12.1)	42-98	172.3 (8.2)	150-186	23.0 (3.5)
Henriksen et al. (17)	50 (26/24)	45.0 (15.5)	19-74	75.3 (16.4)	49-125	173.7 (10.1)	152-193	24.7 (3.6)

Note: Mean (SD) is presented for age, BMI, weight, and height. Ranges and height are rounded to nearest integer. BMI: body mass index, M: male, F: female.

Table 3. Modified COSMIN criteria used for Risk of Bias Assessment.

Parameter and number of studies	Missing Data Report	Missing Data Handling	Sample Size	Acceptable Criterion	Design Flaws	Acceptable Accuracy
Steps (n = 11)	E = 8 (73%) G = 2 (18%) F = 1 (9%) P = 0	E = 8 (73%) G = 2 (18%) F = 0 P = 1 (9%)	E = 0 G = 1 (9%) F = 4 (36%) P = 6 (55%)	E = 9 (82%) G = 0 F = 1 (9%) P = 1 (9%)	E = 5 (46%) G = 4 (36%) F = 2 (18%) P = 0	E = 8 (73%) G = 2 (18%) F = 0 P = 1 (9%)
Physical activity (n = 4)	E = 3 (75%) G = 1 (25%) F = 0 P = 0	E = 3 (75%) G = 1 (25%) F = 0 P = 0	E = 0 G = 1 (25%) F = 1 (25%) P = 2 (50%)	E = 3 (75%) G = 0 F = 0 P = 1 (25%)	E = 1 (25%) G = 1 (25%) F = 2 (50%) P = 0	E = 4 (100%) G = 0 F = 0 P = 0
Energy expenditure (n = 7)	E = 5 (71%) G = 2 (29%) F = 0 P = 0	E = 6 (86%) G = 0 F = 1 (14%) P = 0	E = 0 G = 3 (43%) F = 0 P = 4 (57%)	E = 3 (43%) G = 2 (28.5%) F = 0 P = 2 (28.5%)	E = 3 (43%) G = 1 (14%) F = 3 (43%) P = 0	E = 6 (86%) G = 1 (14%) F = 0 P = 0

Note: E = Excellent, G = Good, F = Fair, P = Poor.

Energy expenditure: Validity of EE was investigated in seven studies (2, 3, 5, 17, 19, 30, 45). Wahl et al. (45), Roos et al. (30), and Boudreaux et al. (3) studied devices in a lab-setting. IC was used as a criterion measure in these three studies using devices previously validated against DLW or other techniques, i.e. Metamax 3B (44), Moxus Modular Metabolic System (31), or TrueOne 2500 (8). Brook et al. (5), Boeselt et al. (2), Hernandez-Vicente et al. (19), and Henriksen et al. (17) conducted their studies under free-living conditions and used accelerometers to estimate EE, using Bodymedia SenseWear armband mini (21), Bodymedia SenseWear (12), ActiGraph ActiTrainer (28), or ActiGraph GT3X (24). MAPE was available in five studies. All studies provided an effect size, either with the ICC, Pearson's correlation coefficient, or both, using accepted activity count cut-offs. An overview of reported correlations is given in Table 4.

When compared to IC in lab-settings, one Polar Loop (45) study showed poor agreement (ICC) for EE, which was generally overestimated when compared to a Metamax 3B. However, in a study under free-living conditions, the Polar Loop (5) was found to have a very strong correlation (Pearson's) when compared to a Bodymedia SenseWear armband mini. MAPE was high in both studies, but lower under free-living conditions. The Polar V800 (30) showed a strong correlation (Pearson's) for EE in lab-settings using IC (Moxus Modular Metabolic System) as criterion measure, but it was significantly underestimated. In contrast, another study compared the Polar V800 (19) with an ActiGraph ActiTrainer under free-living conditions and found a low-to-moderate accuracy, with a significant overestimation of EE. MAPE was high in the lab-based study, and not supplied in the free-living-based study. The Polar A360 (3) showed poor agreement (ICC) for EE in lab-settings testing against IC (TrueOne 2500), with a tendency to overestimate EE. The Polar A300 (2) on the other hand showed good agreement (ICC) compared to a Bodymedia SenseWear when conducted in free-living. Finally, the Polar M430 (17) showed strong correlation and excellent agreement (ICC) when conducted in free-living and compared to an ActiGraph GT3X.

Table 4. MAPE, correlations, and 95% confidence interval for all studies on energy expenditure.

Study	Sub-study	MAPE (%)	Correlation	95% CI
Wahl et al. (45)*	4.3 km/h	56.40*	-0.11†	-0.26-0.31†
	7.2 km/h	53.80*	0.02†	-0.25-0.48†
	10.1 km/h	51.20*	-0.09†	-0.26-0.33†
	13.0 km/h	41.20*	-0.25†	-0.45-0.33†
	Intermittent 10.1, km/h	5.60*	-0.30†	-0.89-0.43†
	Outdoor	22.10*	-0.18†	-0.56-0.48†
	Mean	38.38		
Roos et al. (30)	VO_2 peak 30%	22.76		
	VO_2 peak 50%	11.43		
	VO_2 peak 70%	10.09		
	VO_2 peak 90%	29.98		
	VO_2 peak 110%	39.52		
	Mean	22.76	0.63-0.85^{P}	
Boudreaux et al. (3)		38.18	0.28†	
Brooke et al. (5)		13.00	0.90^{P}	
Boeselt et al. (2)			0.83†	
Hernandez-Vicente (19)	TEE All week		0.48^{P}	
	TEE Weekdays		0.34^{P}	
	TEE Weekend		0.67^{P}	
	TEE-RMR All week		0.57^{P}	
	TEE-RMR Weekdays		0.43^{P}	
	TEE-RMR Weekend		0.74^{P}	
Henriksen et al. (17)	TEE ActiGraph hip triaxial	6.94	0.91^{P}/0.91†	0.78-0.95^{P} 0.80-0.96†
	AEE ActiGraph hip triaxial	24.01	0.76^{P}/0.75†	0.54-0.87^{P} 0.53-0.87†

Note: MAPE: mean absolute percentage error, ICC: intraclass correlation, TEE: total energy expenditure, RMR: resting metabolic rate, AEE: activity energy expenditure, CI: confidence interval. ^{P}Pearson's correlation, †ICC. *Some MAPEs in this study are reported with negative numbers. These MAPES may therefore actually be *mean percentage erro*r and not MAPE.

Physical activity intensity levels: Four studies investigated validity of PA intensity levels (2, 9, 17, 19). One study each used the Polar V800 (Hernandez-Vicente et al. (19)), Polar A300 (Boeselt et al. (2)), Polar M600 (Degroote et al. (9)), and Polar M430 (Henriksen et al. (17)). All studies were conducted under free-living conditions, using different accelerometers previously validated for PA intensity (i.e. ActiGraph ActiTrainer (28), Bodymedia SenseWear (12), or an ActiGraph GT3X (15, 32, 33)). MAPE was reported in one study, but Pearson's correlation coefficients, Spearman's rank correlation coefficient, or ICCs were provided in all studies. An overview of reported correlations is given in Table 5.

The Polar V800 (19) performed well on most tests when compared with an ActiGraph ActiTrainer and gave accurate results for all active time and non-vigorous PA time. However, accuracy was lower in this study for sedentary behavior and vigorous PA. The Polar A300 (2) gave inaccurate results when compared to a Bodymedia SenseWear. The Polar M600 (9) compared to an ActiGraph G3TX showed moderate correlation and poor agreement. The Polar M430 (17) compared to an ActiGraph GT3X was the only PA intensity study reporting MAPE, which was high for all intensities. Strong correlations were found for light PA, vigorous PA, and MVPA. Agreement was mostly poor, with a moderate agreement for vigorous PA.

Table 5. MAPE, correlations, and 95% confidence interval for all studies on physical activity intensity

Study	Sub-study	MAPE (%)	Correlation	95% CI
Boeselt et al. (2)	All days		0.36†	
Hernandez-Vicente et al. (19)	All week			
	Sedentary time		0.69^{P}	
	Active time		0.88^{P}	
	Walking vs lifestyle time		0.52^{P}	
	Walking vs moderate time		0.73^{P}	
	Non-vigorous time		0.85^{P}	
	Vigorous time		0.34^{P}	
	Weekdays			
	Sedentary time		0.66^{P}	
	Active time		0.84^{P}	
	Walking vs lifestyle time		0.49^{P}	
	Walking vs moderate time		0.81^{P}	
	Non-vigorous time		0.81^{P}	
	Vigorous time		0.25^{P}	
	Weekend days			
	Sedentary time		0.76^{P}	
	Active time		0.93^{P}	
	Walking vs lifestyle time		0.57^{P}	
	Walking vs moderate time		0.64^{P}	
	Non-vigorous time		0.90^{P}	
	Vigorous time		0.44^{P}	
Degroote et al. (9)	MVPA, day level		0.53^{s}	0.29-0.72^{s}
			0.38†	0.16-0.56†
	MVPA, 15-minute level		0.15^{s}	0.08-0.41^{s}
			0.46†	0.22-0.51†
Henriksen et al. (17)	Sedentary	29.24	0.52^{P}	0.15-0.73^{P}
			0.33†	0.10-0.51†
	Light	38.09	0.70^{P}	0.53-0.81^{P}
			0.50†	0.37-0.65†
	Moderate	40.89	0.57^{P}	0.27-0.70^{P}
			0.36†	0.18-0.52†
	Vigorous	79.53	0.76^{P}	0.52-0.85^{P}
			0.62†	0.42-0.88†
	MVPA	43.49	0.75^{P}	0.54–0.84^{P}
			0.44†	0.31-0.57†

Note: MAPE: mean absolute percentage error, CI: confidence interval. ^{P}Pearson's correlations, ^{s}Spearman's rank correlation, †ICC.

Step count: Eleven studies investigated step counts validity (1, 2, 6, 9, 14, 17, 19, 37, 38, 45, 46). Bunn et al. (6), Wang et al. (46), Fokkema et al. (14), Smith et al. (38), and Wahl et al. (45) conducted studies in lab-settings counting steps manually. Wahl et al. (45) also used an Optigate system (23). Hernandez-Vicente et al. (19), Boeselt et al. (2), Simunek et al. (37), Degroote et al. (9), and Henriksen et al. (17) conducted studies under free-living conditions, using an ActiGraph ActiTrainer (28), a Bodymedia SenseWear, or an ActiGraph GT3X (22). An et al. (1) used manual step counting in the lab-based sub studies and a New Lifestyle pedometer in the free-living sub-study. Seven studies reported MAPE. All studies reported effect size using Pearson's correlation, Spearman's rank correlation, Cohen's d, Analysis of Variance (ANOVA), and/or ICC. An overview of effect sizes is given in Table 6.

For the studies on Polar Loop (1, 5, 14, 37, 38, 45, 46), agreement varied from poor to good, depending on test setting and walking speed. Some studies reported an overestimation of steps, and some reported an underestimation of steps. In a study conducted under free-living conditions, the Polar A300 (2) gave highly accurate results, when compared to Bodymedia SenseWear, but with a tendency to underestimate steps. The Polar V800 (19), when compared to an ActiGraph, also produced very accurate results, but steps were overestimated. The Polar M600 was only tested in one study (9) where it showed moderate to good agreement with an ActiGraph. Similarly, the Polar M430 (17) was tested against an ActiGraph in one study and was shown to give very strong correlations, moderate agreement, but with high MAPE.

Table 6. MAPE, correlations, and 95% confidence interval for all studies on step counting

Study	Sub-study	MAPE (%)	Correlation/ Cohen's d	95% CI
Bunn et al. (6)	Walking 4.03-6.44 km/h	4.60	0.49^{P}/-0.06‡	-1.18-0.08‡
	Running 8.06-19.3 km/h	10.66	-0.24^{P}/-1.29‡	-0.59--0.94‡
	Mean	7.63		
Wahl et al. (45)	4.3 km/h	-8.70*	0.06†	-0.19-0.40†
	7.2 km/h	-9.60*	-0.27†	-0.55-0.16†
	10.1 km/h	-5.40*	0.39†	-0.08-0.72†
	13.0 km/h	-3.30*	0.69†	0.26-0.88†
	Intermittent, 10.1 km/h	-13.30*	0.31†	0.09-0.70†
	Outdoor	-1.90*	0.83†	0.32-0.96†
	Mean	-7.03		
Wang et al. (46)	Preferred speed w/arm swing	6.75		
	Preferred speed w/arm constraint	10.00		
	Faster speed w/arm swing	6.17		
	Faster speed w/arm constrain	10.94		
	Slower speed w/arm swing	7.87		
	Slower speed w/arm constraint	17.76		
	Walking winding path	7.53		
	Walking on treadmill	11.82		
	Walking up stairs	27.87		
	Walking downstairs	16.27		
	Mean	12.30		
Fokkema et al. (14)	3.2 km/h	24.60	0.08†	-0.15-0.35†
	4.8 km/h	3.00	0.26†	-0.06-0.54†
	6.4 km/h	3.60	0.24†	-0.09-0.53†
	Mean	11.00		
Hernandez-Vicente et al. (19)	All weekdays		0.90^{P}	
	Weekdays		0.89^{P}	
	Weekend days		0.92^{P}	
Boeselt et al. (2)	All speeds		0.99†	
Simunek et al. (37)	ActiGraph	28.0	0.70†	0.17-0.91†
An et al. (1)	Treadmill 3.2 km/h	23.80		
	Treadmill 4.0 km/h	17.60		
	Treadmill 4.8 km/h	15.70		
	Treadmill 5.6 km/h	9.90		
	Treadmill 6.4 km/h	15.20		
	Treadmill 8.0 km/h	14.30		
	Mean	32.17	0.7^{P}	
	Over ground 4 km/h	17.90		
	Over ground 5.2 km/h	17.40		
	Over ground 6.4 km/h	17.80		
	Mean	17.70	0.5^{P}	
	Free-living		0.4^{P}	
Degroote et al. (9)	Day level		0.85^{s}	0.66-0.94^{s}
			0.70†	0.55-0.80†
	15-minute level		0.89^{s}	0.88-0.90^{s}
			0.79†	0.78-0.81†
Henriksen et al. (17)	ActiGraph hip triaxial	25.98	0.85^{P}	0.75-0.91^{P}
			0.63†	0.49-0.75†
Smith et al. (38)	Walking	13.10	0.72†	0.4-0.87†

Note: MAPE: mean absolute percentage error, CI: confidence interval. ^{P}Pearson's correlation, ^{s}Spearman's rank correlation, †ICC, ‡Cohen's d. *Reported with negative numbers in paper and may therefore actually be *mean percentage error* and not MAPE.

DISCUSSION

In this systematic review on Polar wrist-worn devices, we identified 14 validation studies that assessed six different Polar models: Polar Loop (released in 2013), Polar V800 (2014), Polar A300 (2015), Polar A360 (2015), Polar M600 (2016), and Polar M430 (2017). Five Polar models were tested among the seven studies on EE, four studies on PA intensity tested one model each, and five Polar models were tested among the eleven studies on step counting. There were large differences in study setting (i.e. device model, measurement duration, lab vs free-living, and reported metrics), few available studies for each Polar model, and occasionally conflicting result for the same model.
Although there are too few studies to draw a clear conclusion, correlations coefficients (Pearson's/ICC) for EE seem to have increased between studies on earlier models compared to newer models, and studies conducted in free-living reported higher ICC for EE compared to studies conducted in lab-settings. For PA and steps, we could not see any specific trends over time. In addition, although we could not see a clear improvement in MAPE between models over time, there is less error when reporting steps compared to EE and PA. In addition, correlations are occasionally higher for steps compared to EE and PA, but not always.

Step validation for Polar devices generally revealed lower MAPE, and thus higher trustworthiness, compared to other variables, a finding that is in accordance with a systematic review of validation studies on Fitbit devices by Feehan et al. (13). They also found that Fitbit step count accuracy was acceptable in 50% of cases, whereas only two out of 40 Polar sub-studies can be considered to have an acceptable MAPE when using a cut-off of 3% for lab studies (35) and 10% for free-living studies (34). Feehan et al. also found a tendency for Fitbit devices to underestimate steps conducted in lab-settings and overestimate steps in conducted under free-living conditions. This pattern was not clear among currently investigated Polar studies, as some studies overestimated steps and other underestimated steps.

In 2018, Düking et al. (10) published recommendations for studies investigating reliability, sensitivity, and validity when evaluating wearable activity trackers. Regarding validity, they especially underlined the need for an appropriate criterion measure and that all studies should at least calculate the Pearson's correlation coefficient, in order for different studies to be comparable.

The current review identified five studies that reported Pearson's correlation coefficient, six which reported ICC, two which reported both, and one that reported neither. Only about half of the studies that reported Pearson's or ICC also included a confidence interval, limiting the interpretability of the results somewhat. We additionally found that the majority of studies used sufficient criterion measures for steps and PA outcomes, however less than half of the investigated studies used a gold-standard criterion for validation of EE. This result indicates considerable bias and creates difficulty in interpreting and establishing the true ability of Polar wearables to determine human EE.

Furthermore, this review also identified several studies that used an ActiGraph as criterion measure. Even though the ActiGraph is extensively used in research to measure EE, PA intensity, and steps, its accuracy is not yet agreed upon. When compared with indirect calorimetry, studies have found a high correlation for EE and steps when comparing with an Ultima CPX (24), and valid measurements of PA when comparing VO2 measurements by the Cosmed K4b wearable metabolic system (27). However, another study concluded that ActiGraph did not provide valid estimates for EE compared to a portable gas analyzer (MetaMax 3B) (16), and one study found ActiGraph to be valid for step counting only at certain walking speeds when compared to manual step counting (29). Nonetheless, the ActiGraph, and other similar criterions, are currently the best choice when estimating steps and PA under free-living conditions.

The global rise in the use of consumer-based activity trackers have attracted considerable research interest. This is made clear by the increasing amount of validation studies and systematic reviews where researchers are trying to evaluate the potential use for these devices outside the commercial setting. For Polar devices, the observed large differences in study protocols limits interpretability and makes it more difficult to compare results. Researchers who are planning to use consumer-based activity trackers should therefore carefully consider the need for measurement accuracy. If accuracy must be high, an adequately powered pilot test of several activity trackers should preferably be performed before deciding which device to use, as device accuracy seems highly dependent on the study setting. This pilot test setup should be close to the setup of the final study, including participant characteristics.

This is the first systematic review targeting Polar validation studies. The main limitation in this review stems from the large heterogeneity in study protocols and statistical analysis, which made it challenging to evaluate study results. Due to the use of different statistical methods and the lack of available confidence intervals in half of the studies, we could not conduct meta-analysis.

Conclusions: The large differences in study protocols, criterion measures, and statistical analyses, challenge comparisons between devices and concluding how accurate Polar activity trackers are. However, there is less average error for step counting compared to other variables. Future studies should take care to follow guidelines for assessment of wearable sensors designed for physical activity monitoring.

There are too few comparable studies to make a conclusion on free-living versus lab-settings, but the results seem to indicate that the Polar devices are more accurate under free-living, at last when counting steps and measuring EE. Although further studies are required for confirmation, this finding may be of interest since the opposite may be expected due to increased standardization and test leader control in lab-settings. In addition, newer models are now available, and the accuracy of these devices are still unknown. Validation studies on these devices are thus needed.

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. Boeselt T, Spielmanns M, Nell C, Storre JH, Windisch W, Magerhans L, Beutel B, Kenn K, Greulich T, Alter P, Vogelmeier C, Koczulla AR. Validity and usability of physical activity monitoring in patients with chronic obstructive pulmonary disease (copd). PloS one 11(6): e0157229, 2016.

3. Boudreaux BD, Hebert EP, Hollander DB, Williams BM, Cormier CL, Naquin MR, Gillan WW, Gusew EE, Kraemer RR. Validity of wearable activity monitors during cycling and resistance exercise. Med Sci Sports Exerc 50(3): 624-633, 2018.

4. Brickwood K-J, Watson G, O'Brien J, Williams AD. Consumer-based wearable activity trackers increase physical activity participation: Systematic review and meta-analysis. JMIR Mhealth Ueealth 7(4): e11819-e11819, 2019.

5. Brooke SM, An HS, Kang SK, Noble JM, Berg KE, Lee JM. Concurrent validity of wearable activity trackers under free-living conditions. J Strength Cond Res 31(4): 1097-1106, 2017.

6. Bunn JA, Jones C, Oliviera A, Webster MJ. Assessment of step accuracy using the consumer technology association standard. J Sports Sci 37(3): 244-248, 2019.

7. Bunn JA, Navalta JW, Fountaine CJ, Reece JD. Current state of commercial wearable technology in physical activity monitoring 2015-2017. Int J Exerc Sci 11(7): 503-515, 2018.

8. Crouter SE, Antczak A, Hudak JR, DellaValle DM, Haas JD. Accuracy and reliability of the parvomedics trueone 2400 and medgraphics VO2000 metabolic systems. Eur J Appl Physiol 98(2): 139-151, 2006.

9. Degroote L, De Bourdeaudhuij I, Verloigne M, Poppe L, Crombez G. The accuracy of smart devices for measuring physical activity in daily life: Validation study. JMIR Mhealth Uhealth 6(12): e10972, 2018.

10. Duking P, Fuss FK, Holmberg HC, Sperlich B. Recommendations for assessment of the reliability, sensitivity, and validity of data provided by wearable sensors designed for monitoring physical activity. JMIR Mhealth Uhealth 6(4): e102, 2018.

11. 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.

12. Farooqi N, Slinde F, Haglin L, Sandstrom T. Validation of sensewear armband and actiheart monitors for assessments of daily energy expenditure in free-living women with chronic obstructive pulmonary disease. Physiol Rep 1(6): e00150, 2013.

13. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, Hamilton CB, Li LC. Accuracy of fitbit devices: Systematic review and narrative syntheses of quantitative data. JMIR Mhealth Uhealth 6(8): e10527, 2018.

14. Fokkema T, Kooiman TJ, Krijnen WP, Van Der Shans CP, De Groot M. Reliability and validity of ten consumer activity trackers depend on walking speed. Med Sci Sports Exerc 49(4): 793-800, 2017.

15. Freedson PS, Melanson E, Sirard J. Calibration of the computer science and applications, inc. accelerometer. Med Sci Sports Exerc 30(5): 777-781, 1998.

16. Gastin PB, Cayzer C, Dwyer D, Robertson S. Validity of the actigraph gt3x+ and bodymedia sensewear armband to estimate energy expenditure during physical activity and sport. J Sci Med Sport 21(3):291-295, 2018.

17. Henriksen A, Grimsgaard S, Horsch A, Hartvigsen G, Hopstock L. Validity of the polar m430 activity monitor in free-living conditions: Validation study. JMIR Form Res 3(3): e14438, 2019.

18. Henriksen A, Haugen Mikalsen M, Woldaregay AZ, Muzny M, Hartvigsen G, Hopstock LA, Grimsgaard S. Using fitness trackers and smartwatches to measure physical activity in research: Analysis of consumer wrist-worn wearables. J Med Internet Res 20(3): e110, 2018.

19. Hernandez-Vicente A, Santos-Lozano A, De Cocker K, Garatachea N. Validation study of polar v800 accelerometer. Ann Transl Med 4(15): 278, 2016.

20. Hills AP, Mokhtar N, Byrne NM. Assessment of physical activity and energy expenditure: An overview of objective measures. Front Nutr 1:5, 2014.

21. Johannsen DL, Calabro MA, Stewart J, Franke W, Rood JC, Welk GJ. Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. Med Sci Sports Exerc 42(11): 2134-2140, 2010.

22. Lee JA, Williams SM, Brown DD, Laurson KR. Concurrent validation of the actigraph gt3x+, polar active accelerometer, omron hj-720 and yamax digiwalker sw-701 pedometer step counts in lab-based and free-living settings. J Sports Sci 33(10): 991-1000, 2015.

23. Lee M, Song C, Lee K, Shin D, Shin S. Agreement between the spatio-temporal gait parameters from treadmill-based photoelectric cell and the instrumented treadmill system in healthy young adults and stroke patients. Med Sci Monit 20: 1210-1219, 2014.

24. McMinn D, Acharya R, Rowe DA, Gray SR, Allan JL. Measuring activity energy expenditure: Accuracy of the gt3x+ and actiheart monitors. Int J Exerc Sci 6(3): 217-229, 2013.

25. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The cosmin checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international delphi study. Qual Life Res 19(4): 539-549, 2010.

26. Navalta JW, Stone WS, Lyons TS. Ethical issues relating to scientific discovery in exercise science. Int J Exerc Sci 12(1): 1-8, 2019.

27. O'Neil ME, Fragala-Pinkham MA, Forman JL, Trost SG. Measuring reliability and validity of the actigraph gt3x accelerometer for children with cerebral palsy: A feasibility study. J Pediatr Rehabil Med 7(3): 233-240, 2014.

28. Ojiambo R, Konstabel K, Veidebaum T, Reilly J, Verbestel V, Huybrechts I, Sioen I, Casajus JA, Moreno LA, Vicente-Rodriguez G, Bammann K, Tubic BM, Marild S, Westerterp K, Pitsiladis YP, Consortium I. Validity of hip-mounted uniaxial accelerometry with heart-rate monitoring vs. Triaxial accelerometry in the assessment of free-living energy expenditure in young children: The idefics validation study. J Appl Physiol (1985) 113(10): 1530-1536, 2012.

29. Riel H, Rathleff CR, Kalstrup PM, Madsen NK, Pedersen ES, Pape-Haugaard LB, Villumsen M. Comparison between mother, actigraph wgt3x-bt, and a hand tally for measuring steps at various walking speeds under controlled conditions. PeerJ 4: e2799, 2016.

30. Roos L, Taube W, Beeler N, Wyss T. Validity of sports watches when estimating energy expenditure during running. BMC Sports Sci Med Rehabil 9: 22, 2017.

31. Rosdahl H, Lindberg T, Edin F, Nilsson J. The moxus modular metabolic system evaluated with two sensors for ventilation against the douglas bag method. Eur J Appl Physiol 113(5): 1353-1367, 2013.

32. Santos-Lozano A, Santin-Medeiros F, Cardon G, Torres-Luque G, Bailon R, Bergmeir C, Ruiz JR, Lucia A, Garatachea N. Actigraph gt3x: Validation and determination of physical activity intensity cut points. Int J Sports Med 34(11): 975-982, 2013.

33. Sasaki JE, John D, Freedson PS. Validation and comparison of actigraph activity monitors. J Sci Med Sport 14(5): 411-416, 2011.

34. 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.

35. Schneider PL, Crouter SE, Lukajic O, Bassett DR Jr. Accuracy and reliability of 10 pedometers for measuring steps over a 400-m walk. Med Sci Sports Exerc 35(10): 1779-1784, 2003.

36. Shin G, Jarrahi MH, Fei Y, Karami A, Gafinowitz N, Byun A, Lu X. Wearable activity trackers, accuracy, adoption, acceptance and health impact: A systematic literature review. J Biomed Inform 93: 103153, 2019.

37. Simunek A, Dygryn J, Gaba A, Jakubec L, Stelzer J, Chmelik F. Validity of garmin vivofit and polar loop for measuring daily step counts in free-living conditions in adults. Acta Gymnica 46(3): 129-135, 2016.

38. Smith JD, Guerra G, Burkholder BG. The validity and accuracy of wrist-worn activity monitors in lower-limb prosthesis users. Disabil Rehabil doi: 10.1080/09638288.2019.1587792, 2019.

39. Straiton N, Alharbi M, Bauman A, Neubeck L, Gullick J, Bhindi R, Gallagher R. The validity and reliability of consumer-grade activity trackers in older, community-dwelling adults: A systematic review. Maturitas 112: 85-93, 2018.

40. Strath SJ, Kaminsky LA, Ainsworth BE, Ekelund U, Freedson PS, Gary RA, Richardson CR, Smith DT, Swartz AM. Guide to the assessment of physical activity: Clinical and research applications. Circulation 128(20): 2259-2279, 2013.

41. Tang MSS, Moore K, McGavigan A, Clark RA, Ganesan AN. Effectiveness of wearable trackers on physical activity in healthy adults compared: A systematic review and meta-analysis of randomized controlled trials. JMIR Preprints doi: 10.2196/preprints.15576, 2019.

42. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the cosmin checklist. Qual Life Res 21(4): 651-657, 2012.

43. Ubrani J, Llamas R, Shirer M. IDC forecasts sustained double-digit growth for wearable devices led by steady adoption of smartwatches. In: IDC Corporate; 2018.

44. Vogler AJ, Rice AJ, Gore CJ. Validity and reliability of the cortex metamax3b portable metabolic system. J Sports Sci 28(7): 733-742, 2010.

45. 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.

46. Wang L, Liu T, Wang Y, Li Q, Yi J, Inoue Y. Evaluation on step counting performance of wristband activity monitors in daily living environment. IEEE Access 5: 13020-13027, 2017.