©Journal of Sports Science and Medicine (2024) **23,** 79-96
http://www.jssm.org DOI: https://doi.org/10.52082/jssm.2024.79

Research article

Criterion-Related Validity of Consumer-Wearable Activity Trackers for Estimating Steps in Primary Schoolchildren under Controlled Conditions: Fit-Person Study

Daniel Mayorga-Vega [1], Carolina Casado-Robles [2], Santiago Guijarro-Romero [3]✉ and Jesús Viciana [2]
[1] Departamento de Didáctica de las Lenguas, las Artes y el Deporte, Facultad de Ciencias de la Educación, Universidad de Málaga, Málaga, Spain; [2] Department of Physical Education and Sport, University of Granada, Granada, Spain; [3] Department of Didactic of Musical, Plastic and Corporal Expression, University of Valladolid, Valladolid, Spain

Abstract
The purposes were to examine the criterion-related validity of the steps estimated by consumer-wearable activity trackers (wrist-worn activity trackers: Fitbit Ace 2, Garmin Vivofit Jr, and Xiomi Mi Band 5; smartphone applications: Pedometer, Pedometer Pacer Health, and Google Fit/Apple Health) and their comparability in primary schoolchildren under controlled conditions. An initial sample of 66 primary schoolchildren (final sample = 56; 46.4% females), aged 9-12 years old (mean = 10.4 ± 1.0 years), wore three wrist-worn activity trackers (Fitbit Ace 2, Garmin Vivofit Jr 2, and Xiaomi Mi Band 5) on their non-dominant wrist and had three applications in two smartphones (Pedometer, Pedometer Pacer Health, and Google Fit/Apple Health for Android/iOS installed in Samsung Galaxy S20+/iPhone 11 Pro Max) in simulated front trouser pockets. Primary schoolchildren's steps estimated by the consumer-wearable activity trackers and the video-based counting independently by two researchers (gold standard) were recorded while they performed a 200-meter course in slow, normal and brisk pace walking, and running conditions. Results showed that the criterion-related validity of the step scores estimated by the three Samsung applications and the Garmin Vivofit Jr 2 were good-excellent in the four walking/running conditions (e.g., MAPE = 0.6 - 2.3%; lower 95% CI of the ICC = 0.81 - 0.99), as well as being comparable. However, the Apple applications, Fitbit Ace 2, and Xiaomi Mi Band 5 showed poor criterion-related validity and comparability on some walking/running conditions (e.g., lower 95% CI of the ICC < 0.70). Although, as in real life primary schoolchildren also place their smartphones in other parts (e.g., schoolbags, hands or even somewhere away from the body), the criterion-related validity of the Garmin Vivofit Jr 2 potentially would be considerably higher than that of the Samsung applications. The findings of the present study highlight the potential of the Garmin Vivofit Jr 2 for monitoring primary schoolchildren's steps under controlled conditions.

Key words: Validation, wrist-worn activity trackers, smartphone applications, step counts, children, laboratory conditions.

Introduction

Engaging in regular physical activity (PA), especially of moderate-to-vigorous intensity, is widely acknowledged as a significant indicator of health in primary schoolchildren (World Health Organization, 2020). Furthermore, scientific evidence has also shown that total PA is favourably linked to numerous health outcomes in primary schoolchildren (Poitras et al., 2016), with steps per day being a common and reliable measure of total PA (Althof et al., 2017; Craig et al., 2010). The World Health Organization (2020) recommends that primary schoolchildren should engage in at least an average of 60 minutes per day of moderate-to-vigorous PA across the week. However, these PA guidelines are challenging to comprehend for both primary schoolchildren and their parents (Crossley et al., 2019). To address this issue, the moderate-to-vigorous PA-based recommendations have been translated into simpler step-per-day guidelines for primary schoolchildren (Mayorga-Vega et al., 2021). In particular, existing evidence suggests that primary schoolchildren should achieve at least about 10,000 - 12,000 steps per day (Benítez-Porres et al., 2016; Colley et al., 2012; Oliveira et al., 2017).

Consumer-wearable activity trackers have emerged as valuable tools for monitoring and promoting habitual PA among users (Casado-Robles et al., 2022). Such consumer-wearable activity trackers, including wrist-worn activity trackers, clip-on activity trackers and smartphone PA applications, are electronic devices worn on the body to monitor daily PA levels (Casado-Robles et al., 2022). The popularity of consumer-wearable activity trackers has surged in recent years, with global sales of wearable and smartphone devices exceeding 500 million and 13 billion worldwide, respectively (Laricchia, 2023a, 2023b). Given this widespread adoption and their characteristics, stakeholders, including researchers, paediatrics, physical education teachers and parents, are increasingly interested in utilizing consumer-wearable activity trackers to monitor and promote healthy habits of PA in primary schoolchildren (Casado-Robles et al., 2022; Mayorga-Vega et al., 2022).

Among the diverse consumer-wearable activity trackers available, smartphone PA applications and wrist-worn activity trackers have shown to be the most valued and used types of devices by primary schoolchildren (Mayorga-Vega et al., 2022; Viciana et al., 2022). Given that most primary schoolchildren now own smartphones that they carry with them throughout the day (Spanish National Institute of Statistics, 2023) and many PA applications are freely available (Viciana et al., 2022), smartphone PA applications hold a significant advantage as they do not require purchasing any specific device for monitoring and promoting PA. As regards the available purchase options, wrist-worn activity trackers stand out as having several advantages when they are compared with others like clip-on activity trackers, such as reporting real-time feedback that can be easily checked (Maher et al., 2017) or having greater

Received: 23 October 2023 / Accepted: 12 December 2023 / Published (online): 01 March 2024

wear compliance (Fairclough et al., 2016). Moreover, recent scientific evidence supports wrist-worn activity trackers as the most effective for promoting primary schoolchildren's daily PA (Casado-Robles et al., 2022). For these reasons, smartphone PA applications and wrist-worn activity trackers have the potential to serve as feasible tools for objectively monitoring and promoting primary schoolchildren's daily PA (Casado-Robles et al., 2022; Gil-Espinosa et al., 2022; Giurgiu et al., 2022).

Steps per day represent the most common measure for monitoring PA and personalized goal-setting for promoting PA through consumer-wearable activity trackers (Casado-Robles et al., 2022; Maher et al., 2017). However, before utilizing a particular consumer-wearable activity tracker, it is crucial to assess its validity and ensure its appropriateness for the target population (Kottner et al., 2011; Mokkink et al., 2010). Criterion-related validity of step counts estimated by consumer-wearable activity trackers should be analyzed by examining the agreement between their scores and those from the "gold standard", which currently involves video-based counting conducted by at least two observers (Johnston et al., 2021). The best-practice protocol for the validation of steps estimated by consumer-wearable activity trackers should be conducted under controlled, semi-free living, and free-living conditions (Johnston et al., 2021). The controlled testing condition, which involves participants wearing the activity trackers while completing walking/running tasks at controlled or self-selected speeds, represents the first stage in the multistage protocols for the best-practice validation of steps estimated by consumer-wearable activity trackers (Johnston et al., 2021). Furthermore, since different kinds of consumer-wearable activity trackers could be used in the same context due to economic constrains (e.g., monitoring or promoting PA in the Physical Education setting or large-scale research studies) (Brodie et al., 2018; Creaser et al., 2022), the agreement between different devices (i.e., comparability) should be also studied (Viciana et al., 2022).

In spite of the increasing use of smartphone PA applications and wrist-worn activity trackers, there is a lack of substantial evidence regarding their criterion-related validity and comparability in primary schoolchildren. To date, and to our knowledge, only two prior studies have examined the criterion-related validity of steps estimated by wrist-worn activity trackers in primary schoolchildren under controlled conditions (Godino et al., 2020; Sun et al., 2022). These studies found that the wrist-worn activity trackers Fitbit Charge HR (Godino et al., 2020), Fitbit Ace, and Moki (Sun et al., 2022) had good to excellent criterion-related validity for estimating steps. Moreover, as far as we know, no previous topic-related studies were carried out with smartphone PA applications among primary schoolchildren. Furthermore, to the best of our knowledge, there is a lack of prior studies examining the comparability of steps estimated by smartphone PA applications and wrist-worn activity trackers in this population.

Consequently, the main purpose of the present study was to examine the criterion-related validity of the steps estimated by the consumer-wearable activity trackers (wrist-worn activity trackers: Fitbit Ace 2, Garmin Vivofit Jr, and Xiomi Mi Band 5; smartphones applications: Pedometer, Pedometer Pacer Health, and Google Fit/Apple Health) in primary schoolchildren under controlled conditions. The secondary purpose of this study was to examine the comparability of the steps estimated by the above-mentioned consumer-wearable activity trackers in primary schoolchildren under controlled conditions.

Methods

Participants

The present study is reported according to the GRRAS guidelines (Kottner et al., 2011). The protocol of the present study conforms to the Declaration of Helsinki statements (64th WMA, Brazil, October 2013) and it was first approved by the Ethical Committee for Human Studies at the University of Granada. Three public primary schools located in urban areas of the province of Granada (Spain) chosen by convenience. According to the schools' reports, all the primary schoolchildren's families had a middle socioeconomic level. The principal and the PE teachers were first contacted. Then, they were informed about the project, and permission to conduct the study was requested. After the approvals of the schools was obtained, all the primary schoolchildren and their legal guardians were fully informed about the features of the project. Primary schoolchildren's verbal informed assents and their legal guardians' signed written informed consents were obtained before taking part in the study.

The present study followed a cross-sectional design. A total of 66 primary schoolchildren from 4th to 6th grade (i.e., 9 - 12 years old) enrolled in the selected schools were invited to participate in the present study. The following inclusion criteria were considered: a) being enrolled in the 4th to 6th grade at the primary education level (i.e., target grades according to study aim); b) being free of any health disorder that would make them unable to engage in PA normally; c) providing the corresponding verbal informed assents of the primary schoolchildren, and d) presenting the corresponding signed written informed consents of the primary schoolchildren's legal guardians. The following exclusion criteria were considered: a) not having completed and valid data from the five wearable activity trackers, and/or b) not having completed and valid data from the video-based step count.

A priori sample size calculation was estimated with the Arifin's web-based sample size calculator (Arifin, 2018). Parameters were set as follows: ICC, $\rho_0 = 0.70$ (Nunnally, 1978); $\rho_1 = 0.85$ (Viciana et al., 2022), $\alpha = 0.05$, $1\text{-}\beta = 0.80$, $k = 2$, dropout = 10% (Viciana et al., 2022). A final sample size of at least 53 primary schoolchildren (minimum initial sample size = 59) was estimated. In addition to exceeding the minimum required sample size, the aim for each study sampling was to obtain a sample balanced by grade and gender.

Measures

Demographic characteristics. Primary schoolchildren's grade ($4^{th}/5^{th}/6^{th}$), gender (males/females), age (in years) and non-dominant hand (left/right) information was self-reported in a written questionnaire.

Anthropometric. Primary schoolchildren's body

mass (kg) and height (cm) were first measured following the International Standards for Anthropometric Assessment (Stewart et al., 2011). Firstly, primary schoolchildren's body mass and height were measured in shorts, T-shirts, and barefoot. For the body mass measure, primary schoolchildren stood in the centre of the scale (Seca, Ltd., Hamburg, Germany; accuracy = 0.1 kg) without support and with the weight distributed evenly on both feet. For the body height assessment, primary schoolchildren stood with their feet together with the heels, buttocks and upper part of the back touching the stadiometer (Holtain Ltd., Crymmych, Pembs, United Kingdom; accuracy = 0.1 cm), and with the head placed in the Frankfort plane. Each measurement was performed twice and the mean was recorded (Stewart et al., 2011). Then, the body mass index was calculated as body mass divided by body height squared (kg/m2). Finally, primary schoolchildren's body weight status was categorized by gender- and age-adjusted body mass index thresholds as overweight/obesity or non-overweight/obesity (Cole et al., 2000). Body mass index and body weight status scores have shown high evidence supporting validity for body composition among primary schoolchildren (Cole et al., 2000).

Consumer-wearable activity trackers. Primary schoolchildren's steps were estimated by three wrist-worn activity trackers [Fitbit Ace 2 (Fitbit, San Francisco, SF, USA), Garmin Vivofit Jr 2 (Garmin, Kansas, KS, USA), and Xiaomi Mi Band 5 (Xiaomi, Pekin, China)] and three applications in two smartphones [Pedometer (ITO Technologies) and Pedometer Pacer Health for Android (Samsung Galaxy S20+) and iOS (iPhone 11 Pro Max); and Google Fit application for Android (Samsung Galaxy S20+), and the Apple Health application for iOS (iPhone 11 Pro Max)]. Physical specifications of the chosen devices are as follows: Fitbit Ace 2: 2.27 x 1.00 x 0.30 cm, 20.0 g; Garmin Vivofit Jr 2: 1.1 x 1.1 x 0.9 cm, 17.5 g; Xiaomi Mi Band 5: 4.69 x 1.81 x 1.24 cm, 11.9 g; Samsung Galaxy S20+: 16.2 x 7.4 x 0.8 cm, 186 g, and iPhone 11 Pro Max: 15.8 x 7.8 x 0.8 cm, 226 g. The three chosen wrist-worn activity trackers are based in tri-axial built-in accelerometers, while the chosen smartphones have different sensors including accelerometers and gyroscopes. Each device and application has its own proprietary algorithm to estimate the step counts.

Concerning the particular chosen activity trackers, the criteria were as follows: a) the most worldwide used display-based activity wristbands brands (Henriksen et al., 2018; IDC's Worldwide Quarterly Wearable Device Tracker reports from 2017 to 2020); b) choosing models of the devices with affordable prices (based on launch prices in Spain; Fitbit Ace 2 ≈ 70€; Garmin Vivofit Jr 2 ≈ 70€; Xiaomi Mi Band 5 ≈ 35€); c) choosing the most advanced model (in that moment), and d) models designed specifically for children, when they were available (i.e., Garmin Vivofit Jr 2 and Fitbit Ace 2). For the smartphone applications, the criteria were to study: a) applications for Android and iOS, and b) choosing the most popular and used free downloadable applications available in the applications stores (due to the number of downloads and their user ratings) and the included applications of the corresponding smartphones (i.e., Samsung Google Fit for Android and Apple Health for iOS). As regards the specific smartphones used, the criteria were the most worldwide used brands (IDC's Worldwide Quarterly Wearable Device Tracker reports from 2017 to 2020) and choosing the most advanced model (in that moment) for Android and iOS.

Finally, as regards the number of wrist-worn activity trackers, it was considered that three wrist-worn activity trackers and two smartphones were a feasible number that did not interfere with the primary schoolchildren's movements while walking and running (i.e., natural arm and leg swing) and allowed for a correct measurement (i.e., wrist and legs adjustment). In this line, the total mass of the three wrist-worn activity trackers (37.5 g) and two smartphones (186 or 226 g in each thigh) was not high. According to the user manual of the wrist-worn activity trackers, one device of each model was adjusted snugly on the top of primary schoolchildren's non-dominant wrist, close to and above the wrist bone (they were 3.91 cm width). Regarding the smartphones, one device of each model was allocated in two bags (i.e., one in each bag), adjusted snugly with a belt, on the top and front part of the primary schoolchildren's thighs (one in each) as if they were placed in trouser pockets and did not interfere with the primary schoolchildren's movements during the trials. Activity trackers were adjusted so they could not move, but overtightening was avoided.

Video-based steps count. Primary schoolchildren's steps gold standard was determined by step counting the video recording in slow-motion (Johnston et al., 2021). Primary schoolchildren were asked to perform a 200-meter course in four different conditions. The 200-meter course was marked with cones and lines and performed inside the school on a non-slippery sport court with an oval shape and no tight turns. A digital video camera (Go Pro Hero 7, California, USA) with a tripod was situated in the middle of the sports court in order to easily record the primary schoolchildren's lower limbs during the entire course from the sagittal plane. For calculating the speed and step cadence of each condition, time was considered as from when the primary schoolchildren started walking/running until they crossed the finish line. The gold standard step count for each schoolchild in each condition was performed independently by two researchers through the slow-motion video recording projected on a 15.6" screen. When disagreement occurred (8.6%), these particular observations were evaluated again by the two researchers. Although most of the disagreements were simply due to an error in one of the two researchers, when disagreement still occurred, a third researcher evaluated it.

Procedure

Evaluations were carried out during the afternoon in participants' leisure time from Monday to Friday, and then data were downloaded and batteries charged during the morning. Due to the limitations of material and human resources, about two or three primary schoolchildren per hour were evaluated one by one during each evaluation session. Data collection was carried out by the same researchers, instruments and protocols. Firstly, primary schoolchildren's demographic characteristics and anthropometric measurements were recorded. Then, the five devices were

adjusted on primary schoolchildren. In order to avoid the relative position of the activity trackers influencing the outcomes, they were adjusted in a random order varying across the primary schoolchildren (i.e., the position on the non-dominant wrist from hand to elbow for the wrist-worn activity trackers and the left/right thigh for the smartphones) (Hartung et al., 2020).

Finally, primary schoolchildren were instructed to walk/run the 200-meter course in the following four conditions, at a continuous speed, and with a natural arm and leg swing: 1) slow pace walking; 2) normal pace walking (self-pace walking); 3) brisk pace walking; and 4) running (jogging). Participants chose their walking/running speed based on the instructions provided for each condition (e.g., for the normal pace walking condition: "Perform the course at a speed that corresponds to walking naturally, at an everyday walking pace. For example, similar to the one you follow when going from home to school"). Before starting, a demonstration in order to guide each participant was performed. When primary schoolchildren were at the starting line, the steps count from the activity trackers was recorded. Then, they were instructed to not move until they started walking/running. They also were asked to always start the course with the contralateral leg to the arm where the wrist-worn activity trackers were attached. Primary schoolchildren were requested to stop immediately after the finish line, and a cone was situated five meters beforehand to remind them. Then, the steps counted by the activity trackers were registered.

Statistical analysis

Descriptive statistics for all the variables of the included participants were calculated. Firstly, all the statistical tests assumptions were checked (e.g., histograms and Q-Q plots for normality) and met. Furthermore, univariate (i.e., z ± 3.0) and multivariate outliers (i.e., Mahalanobis distance) were removed. Afterward, for examining the main purpose of the present study (i.e., criterion-related validity), the agreement between the number of steps assessed by the consumer-wearable activity trackers and the video-based count (gold standard) were calculated as follows: a) Equivalence test with the 90% confidence interval (CI) method (Dixon et al., 2018); b) Limits of Agreement (LOA) with its 95% CI (Bland and Altman, 1986); c) Mean Absolute Error (MAE) (Willmott and Matsuura, 2005); d) Mean Absolute Percentage Error (MAPE) (Johnston et al., 2021), and e) Intraclass Correlation Coefficient (ICC), and its 95% CI, by a two-way random effects model with absolute agreement and single measurement [also known as ICC(2,1)] (Koo and Li, 2016). Based on previous literature, agreement values were interpreted as follows: Equivalence test, when the mean reference standard score is within ± 15% of the mean consumer-wearable activity trackers score is considered acceptable (Dixon et al., 2018); MAPE, > 15.0% poor, 10.1 - 15.0% acceptable, 5.1 - 10.0% good, and 0.0 - 5.0% excellent (Johnston et al., 2021); ICC, 0.00 - 0.69 poor, 0.70 - 0.79 acceptable, 0.80 - 0.89 good, and 0.90 - 1.00 excellent (Nunnally, 1978). Based on statistical inference, each ICC value was interpreted according to its 95% CI, that means, there was a 95% chance that the true ICC value landed on any point between the 95% CI range (Koo and Li, 2016). Finally, LOA plots, which are the individual participant differences between the two scores plotted against the respective individual means, were performed (Bland and Altman, 1995). Heteroscedasticity was also examined objectively by calculating the Pearson's correlation coefficient (r) between the absolute differences and the individual means (Atkinson and Nevill, 1998). Based on Cohen's (Cohen, 1992) benchmarks, a correlation coefficient > 0.50 was considered as indicative of heteroscedasticity. Finally, as regards the secondary purpose of the present study (i.e., comparability), similarly the agreement between the number of steps estimated by pairs of consumer-wearable activity trackers was examined. All statistical analyses were performed using the SPSS version 25.0 for Windows (IBM® SPSS® Statistics), except for the equivalence test where the Jamovi version 2.3 (The Jamovi project, https://www.jamovi.org) was used. The statistical significance level was set at $p < 0.05$.

Results

General characteristics

Figure 1 shows the flow diagram of the participants throughout the study. From the 66 primary schoolchildren that were invited to participate in the present study, 63 primary schoolchildren agreed and met the inclusion criteria. Since some primary schoolchildren met at least one exclusion criterion, the final sample consisted of 56 participants (i.e., non-compliance rate of 11.1%). Table 1 shows the general characteristics of the participants.

Table 1. General characteristics of the participants

	Eligible sample (*n* = 63)	Final sample (*n* = 56)
Age (years)[a]	10.4 (1.0)	10.4 (0.9)
Grade (4th/5th/6th)[b]	36.5/31.7/31.7	35.7/32.1/32.1
Gender (males/females)[b]	50.8/49.2	53.6/46.4
Body mass (kg)[a]	38.9 (8.4)	39.2 (8.0)
Body height (cm)[a]	143.9 (7.3)	144.4 (7.2)
Body mass index (kg/m2)[a]	18.6 (3.0)	18.7 (2.9)
Overweight/obesity (no/yes)[b]	77.8/22.2	78.6/21.4
Non-dominant hand (left/right)[b]	93.7/6.3	94.6/5.4

Data are reported as mean (standard deviation) [a] or percentage [b]. PA = Physical activity.

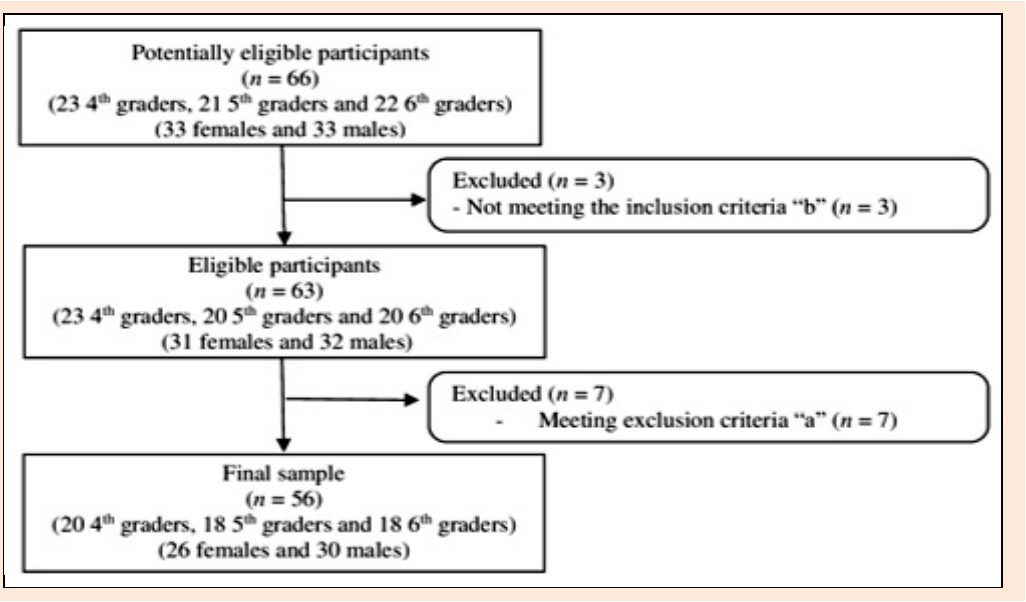

Figure 1. **Flow diagram of the participants throughout the study.**

Table 2. **Criterion-related validity of the consumer-wearable activity trackers for estimating steps during controlled conditions (*n* = 56).**

Instrument	Mean (SD)	Equivalence test (90% CI)	LOA (95% CI)	MAE	MAPE	ICC (95% CI)
Slow pace walking						
Video-based count	335.3 (28.8)	-50.30, 50.30	-	-	-	-
Samsung Pedometer	334.8 (29.4)	-0.54, 1.65	0.6 (-9.0, 10.2)	3.3	1.0	0.99 (0.98, 0.99)
Samsung Pacer	334.7 (29.2)	-0.38, 1.59	0.6 (-8.0, 9.2)	2.9	0.9	0.99 (0.98, 0.99)
Samsung GoogleFit	334.7 (29.2)	-0.33, 1.58	0.6 (-7.8, 9.0)	2.8	0.9	0.99 (0.98, 0.99)
Apple applications [a]	332.6 (27.8)	0.81, 4.59	2.7 (-14.0, 19.4)	3.9	1.1	0.95 (0.92, 0.97)
Fitbit Ace 2	329.6 (30.4)	2.81, 8.55	5.7 (-19.4, 30.8)	8.2	2.5	0.89 (0.79, 0.94)
Garmin Vivofit Jr 2	334.1 (28.4)	0.02, 2.44	1.2 (-9.4, 11.8)	3.5	1.0	0.98 (0.97, 0.99)
Xiaomi Mi Band 5	334.0 (29.8)	0.17, 2.47	1.3 (-8.9, 11.5)	3.3	1.0	0.98 (0.97, 0.99)
Normal pace walking						
Video-based count	296.8 (23.2)	-44.52, 44.52	-	-	-	-
Samsung Pedometer	297.8 (23.9)	-2.14, 0.24	-0.9 (-11.3, 9.5)	3.8	1.3	0.97 (0.96, 0.99)
Samsung Pacer	296.0 (22.8)	0.29, 1.32	0.8 (-3.7, 5.3)	1.7	0.6	0.99 (0.99, 1.00)
Samsung GoogleFit	295.8 (23.1)	0.07, 1.96	1.0 (-7.2, 9.2)	2.2	0.7	0.98 (0.97, 0.99)
Apple applications [a]	294.7 (22.5)	0.86, 3.39	2.1 (-9.1, 13.3)	2.7	0.9	0.97 (0.94, 0.98)
Fitbit Ace 2	286.8 (26.8)	7.36, 12.67	10.0 (-13.3, 33.3)	10.9	3.7	0.82 (0.42, 0.93)
Garmin Vivofit Jr 2	297.9 (24.8)	-2.69, 0.61	-1.0 (-15.5, 13.5)	4.9	1.6	0.95 (0.92, 0.97)
Xiaomi Mi Band 5	286.6 (31.3)	6.40, 14.14	10.3 (-23.6, 44.2)	11.7	4.0	0.75 (0.50, 0.87)
Brisk pace walking						
Video-based count	259.8 (17.8)	-38.97, 38.97	-	-	-	-
Samsung Pedometer	260.2 (17.8)	-1.56, 0.81	-0.4 (-10.8, 10.0)	3.7	1.5	0.96 (0.93, 0.97)
Samsung Pacer	259.1 (17.7)	0.07, 1.39	0.7 (-5.2, 6.6)	2.1	0.8	0.99 (0.98, 0.99)
Samsung GoogleFit	258.4 (18.5)	0.07, 2.78	1.4 (-10.6, 13.4)	2.8	1.1	0.94 (0.90, 0.97)
Apple applications [a]	248.3 (22.2)	7.18, 15.82	11.5 (-26.3, 49.3)	12.0	4.5	0.47 (0.17, 0.67)
Fitbit Ace 2	236.8 (26.2)	18.68, 27.28	23.0 (-14.6, 60.6)	23.8	9.2	0.42 (0.00, 0.71)
Garmin Vivofit Jr 2	259.2 (18.0)	-1.30, 2.58	0.6 (-16.5, 17.7)	5.9	2.3	0.88 (0.81, 0.93)
Xiaomi Mi Band 5	243.3 (28.8)	10.95, 22.16	16.6 (-32.6, 65.8)	17.9	6.9	0.37 (0.07, 0.60)
Running						
Video-based count	223.1 (32.0)	-33.47, 33.47	-	-	-	-
Samsung Pedometer	221.6 (32.0)	0.64, 2.36	1.5 (-5.9, 8.9)	3.0	1.4	0.99 (0.98, 1.00)
Samsung Pacer	221.0 (31.3)	0.43, 3.64	2.0 (-12.1, 16.1)	3.6	1.5	0.97 (0.95, 0.98)
Samsung GoogleFit	221.4 (31.0)	0.00, 3.39	1.7 (-13.2, 16.6)	3.9	1.7	0.97 (0.95, 0.98)
Apple applications [a]	220.7 (30.9)	0.73, 4.02	2.4 (-12.1, 16.9)	3.3	1.4	0.97 (0.95, 0.98)
Fitbit Ace 2	209.8 (42.2)	8.81, 17.69	13.3 (-25.7, 52.3)	14.8	7.1	0.81 (0.54, 0.91)
Garmin Vivofit Jr 2	223.6 (32.3)	-1.51, 0.51	-0.5 (-9.3, 8.3)	3.3	1.5	0.99 (0.98, 0.99)
Xiaomi Mi Band 5	220.4 (32.1)	1.33, 4.10	2.7 (-9.5, 14.9)	4.3	2.0	0.98 (0.96, 0.99)

SD = Standard deviation; LOA = Limits of Agreement; 90/95% CI = 90/95% Confident Interval; MAE = Mean Absolute Error; MAPE = Mean Absolute Percentage Error; ICC = Intraclass Correlation Coefficient. [a] Apple applications is referred to the three applications activated in the iPhone smartphone (i.e., Pedometer, Pacer, and Apple Health) due to the fact that all of them reported exactly the same steps scores.

Criterion-related validity of the consumer-wearable activity trackers for estimating steps

Table 2 shows the criterion-related validity of the consumer-wearable activity trackers for estimating steps during controlled conditions. The results showed that the criterion-related validity of the step scores estimated by the activity trackers tended to be higher for slow pace walking, followed by running, normal pace walking and brisk pace walking. Particularly, the results showed that the criterion-related validity of the step scores estimated by the three Samsung applications were excellent in all of the four walking/running conditions (e.g., scores inside the 90% CI of the equivalence test, MAPE ≤ 5%, and 95% CI of the ICC ≥ 0.90). Similarly, the criterion-related validity results of the steps estimated by the Garmin Vivofit Jr 2 was excellent, except for the 95% CI of the ICC value on the brisk pace walking condition, that was good.

However, regarding the Apple applications, although most of the criterion-related validity results were excellent, the 95% CI of the ICC value on the brisk pace walking condition was poor (since in the iPhone 11 Pro Max the three applications reported exactly the same steps scores, note that results are reported as "Apple applications"). Moreover, although most of the criterion-related validity results of the steps estimated by the Xiaomi Mi Band 5 ranged from good to excellent, the 95% CI of the ICC values for the normal and brisk pace walking conditions were poor. Furthermore, the criterion-related validity results of the steps estimated by the Fitbit Ace 2 with the 95% CI of the ICC ranged from poor to acceptable (but scores inside the 90% CI of the equivalence test, and MAPE values ranged from good to excellent).

Figure 2, Figure 3, Figure 4 and Figure 5 shows the LOA plots. Pearson's correlation coefficients did not show heteroscedasticity on any walking/running condition (r = -0.49-0.21), except with the Fitbit Ace 2 on the running condition (r = - 0.52; Table 3). The average speed (SD) in each condition was as follows: Slow pace walking = 1.1 (0.1) m/s [4.0 (0.5) km/h]; normal pace walking = 1.4 (0.1) m/s [5.0 (0.5) km/h]; brisk pace walking = 1.8 (0.2) m/s [6.4 (0.5) km/h]; and running = 2.7 (0.4) m/s [9.7 (1.3) km/h]. The average steps cadence (SD) with the video-based count in each condition was as follows: Slow pace walking = 109.8 (8.3) steps/min; normal pace walking = 122.5 (7.5) steps/min; brisk pace walking = 138.7 (8.6) steps/min; and running = 177.9 (10.8) steps/min.

Table 3. Pearson's correlation coefficient (r) between the absolute differences and the individual means (n = 56).

Instrument	Slow pace walking (steps)	Normal pace walking (steps)	Brisk pace walking (steps)	Running (steps)
Criterion-related validity				
Samsung Pedometer	0.08	0.16	-0.07	-0.07
Samsung Pacer	-0.04	-0.11	0.03	0.15
Samsung GoogleFit	-0.04	-0.02	-0.13	0.09
Apple applications [a]	0.10	0.16	-0.21	0.20
Fitbit Ace 2	-0.16	-0.30*	-0.47†	-0.52†
Garmin Vivofit Jr 2	-0.08	0.21	-0.19	-0.02
Xiaomi Mi Band 5	-0.03	-0.45‡	-0.49†	0.00
Comparability				
Samsung Pedometer - Samsung Pacer	0.26	0.10	0.03	0.22
Samsung Pedometer - Samsung GoogleFit	0.26	0.11	-0.11	0.18
Samsung Pedometer - Apple applications [a]	0.14	0.23	-0.22	0.11
Samsung Pedometer - Fitbit Ace 2	-0.05	-0.24	-0.47†	-0.51†
Samsung Pedometer - Garmin Vivofit Jr 2	0.11	0.19	-0.18	-0.08
Samsung Pedometer - Xiaomi Mi Band 5	0.17	-0.41‡	-0.48†	-0.01
Samsung Pacer - Samsung GoogleFit	0.01	0.08	-0.18	-0.06
Samsung Pacer - Apple applications [a]	0.07	0.11	-0.21	0.21
Samsung Pacer - Fitbit Ace 2	-0.11	-0.31*	-0.45‡	-0.50†
Samsung Pacer - Garmin Vivofit Jr 2	-0.10	0.19	-0.13	0.14
Samsung Pacer - Xiaomi Mi Band 5	0.12	-0.45†	-0.47†	0.15
Samsung GoogleFit - Apple applications [a]	0.07	0.10	-0.24	0.10
Samsung GoogleFit - Fitbit Ace 2	-0.12	-0.29*	-0.41‡	-0.51†
Samsung GoogleFit - Garmin Vivofit Jr 2	-0.12	0.14	-0.21	0.11
Samsung GoogleFit - Xiaomi Mi Band 5	0.10	-0.44‡	-0.47†	0.11
Apple applicationsa - Fitbit Ace 2	-0.18	-0.19	-0.39‡	-0.52†
Apple applicationsa - Garmin Vivofit Jr 2	0.15	0.23	-0.25	0.19
Apple applicationsa - Xiaomi Mi Band 5	0.21	-0.35‡	-0.42‡	0.26
Fitbit Ace 2 - Garmin Vivofit Jr 2	-0.11	-0.23	-0.45‡	-0.49†
Fitbit Ace 2 - Xiaomi Mi Band 5	-0.07	-0.31*	-0.19	-0.45‡
Garmin Vivofit Jr 2 - Xiaomi Mi Band 5	0.20	-0.41‡	-0.53†	0.09

[a] Apple applications is referred to the three applications activated in the iPhone smartphone (i.e., Pedometer, Pacer, and Apple Health) due to the fact that all of them reported exactly the same steps scores. * $p < 0.05$, ‡ $p < 0.01$, and † $p < 0.001$

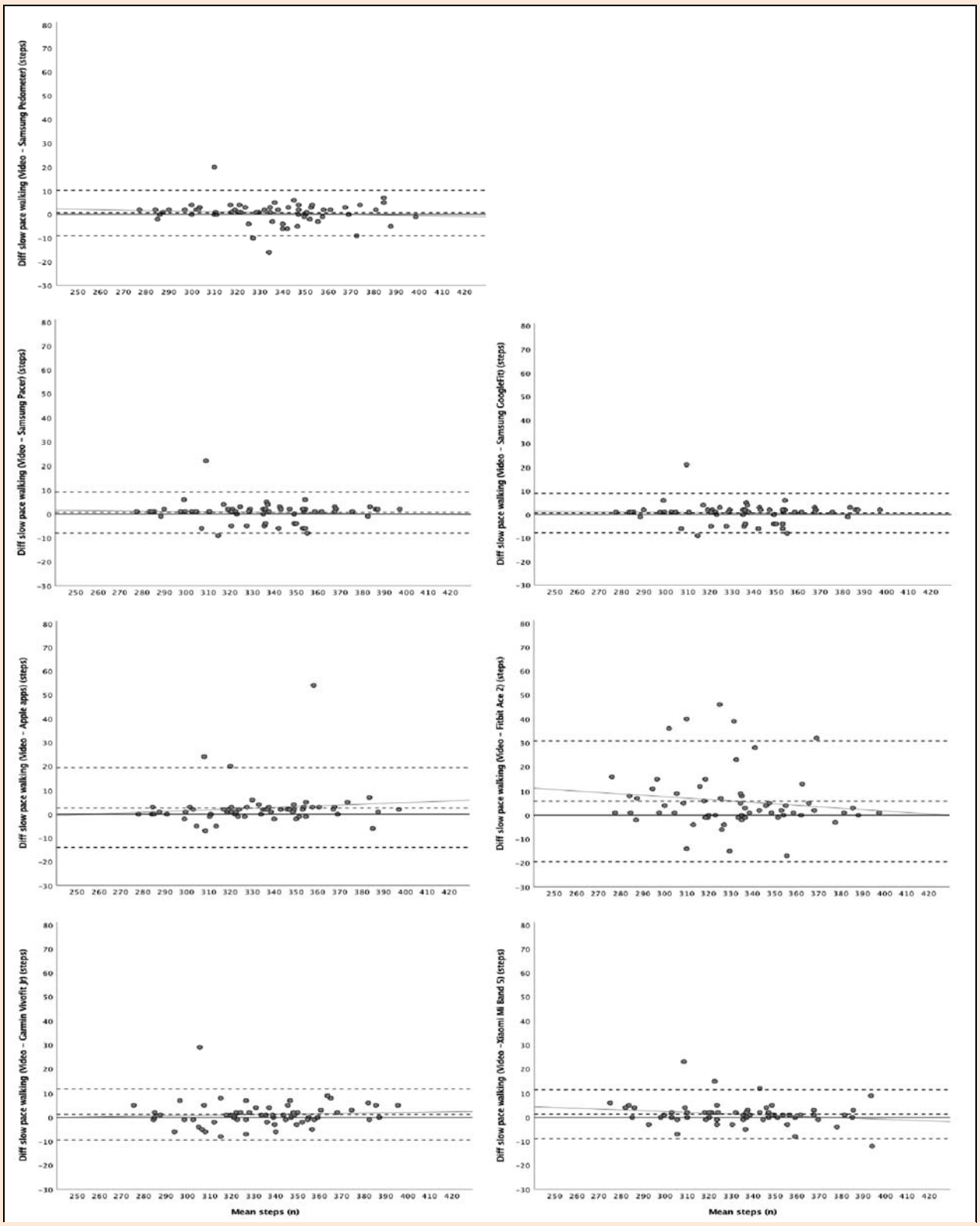

Figure 2. Limits of agreement plots of the consumer-wearable activity trackers for estimating steps during controlled conditions (slow pace walking condition). The middle-dashed line indicates the mean difference (systematic bias) between step scores assessed by the consumer-wearable activity trackers and the video-based count (gold standard) and the upper and lower dashed lines indicate the limits of agreement (95% confidence interval).

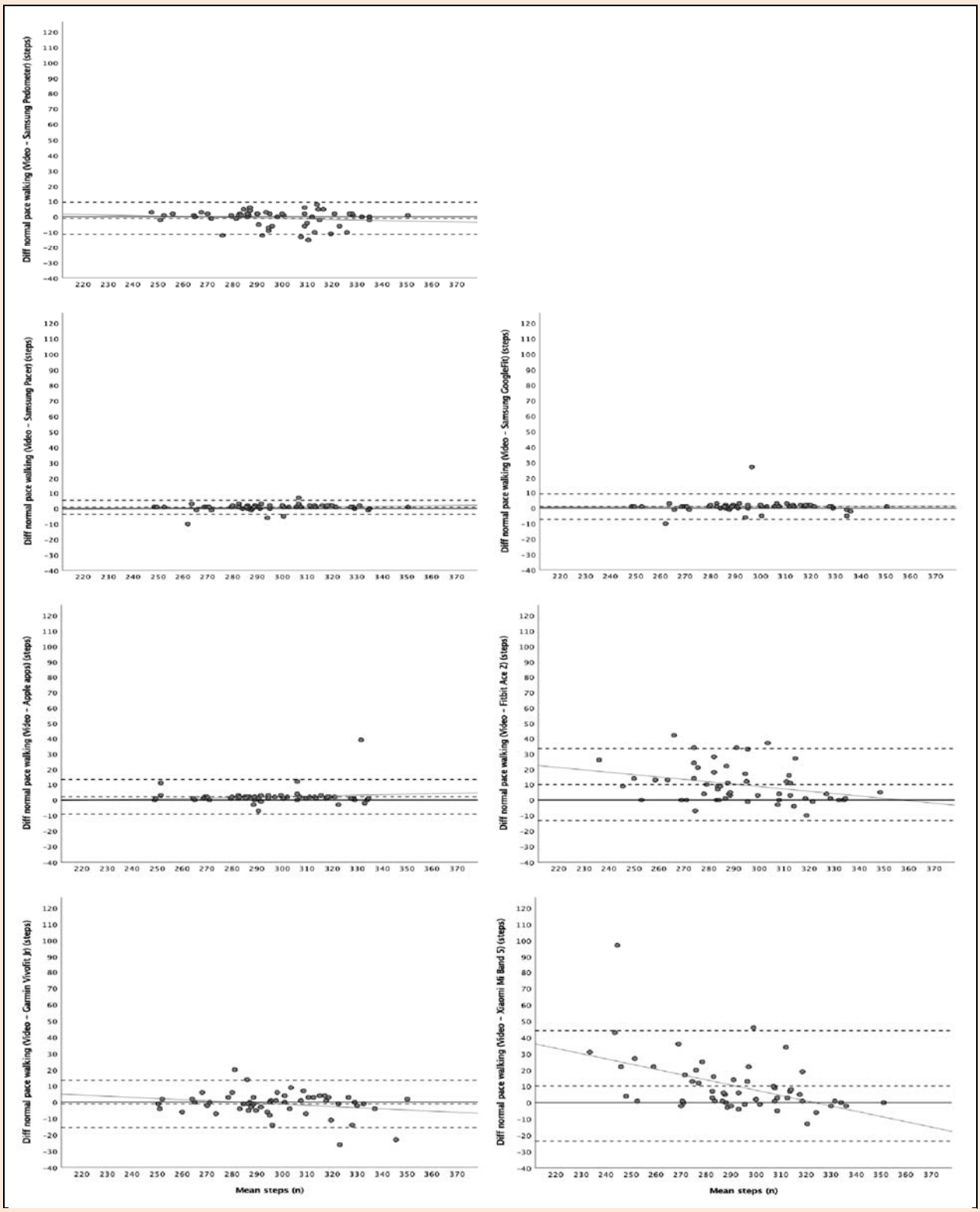

Figure 3. Limits of agreement plots of the consumer-wearable activity trackers for estimating steps during controlled conditions (normal pace walking condition). The middle-dashed line indicates the mean difference (systematic bias) between step scores assessed by the consumer-wearable activity trackers and the video-based count (gold standard) and the upper and lower dashed lines indicate the limits of agreement (95% confidence interval).

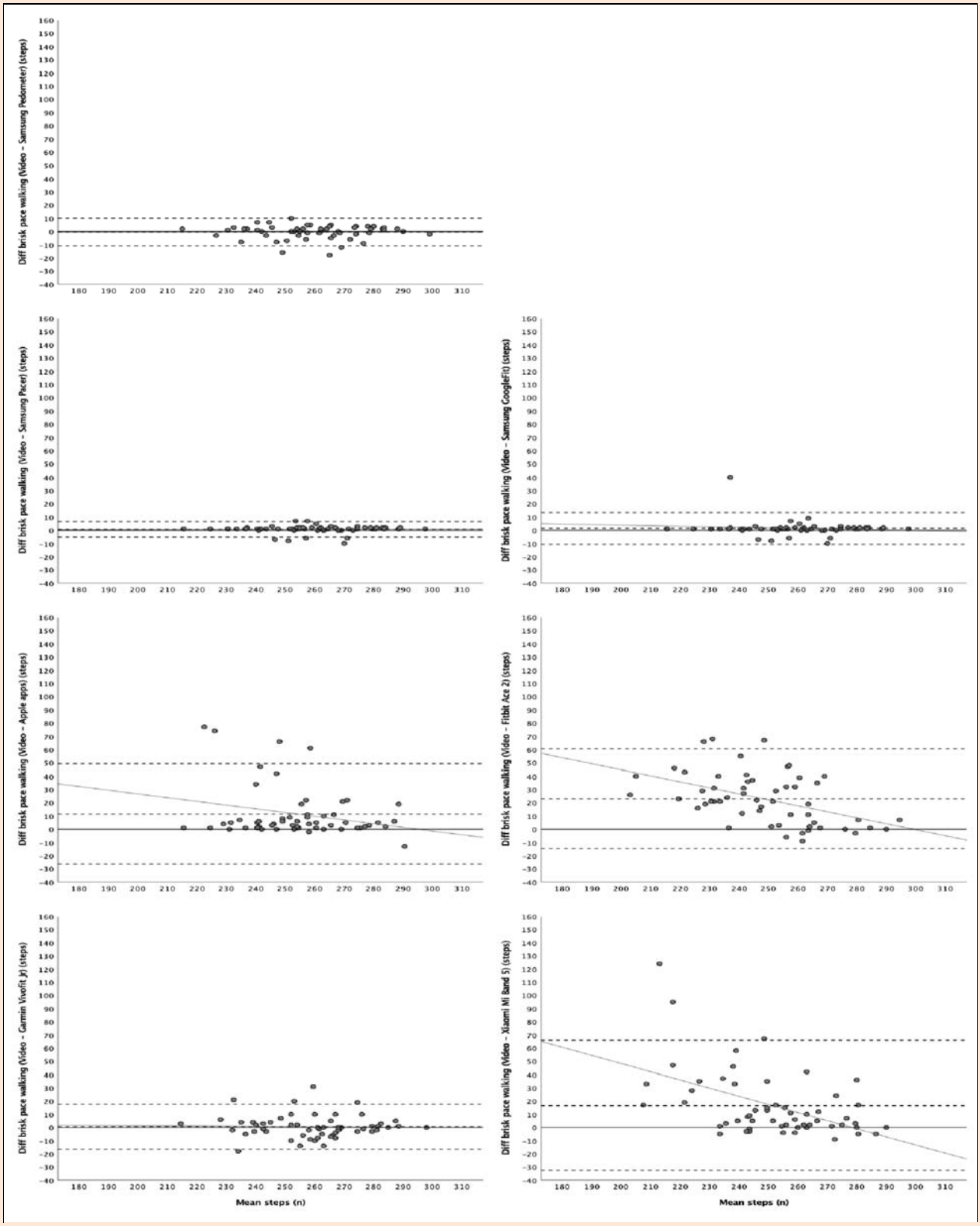

Figure 4. Limits of agreement plots of the consumer-wearable activity trackers for estimating steps during controlled conditions (brisk pace walking condition). The middle-dashed line indicates the mean difference (systematic bias) between step scores assessed by the consumer-wearable activity trackers and the video-based count (gold standard) and the upper and lower dashed lines indicate the limits of agreement (95% confidence interval).

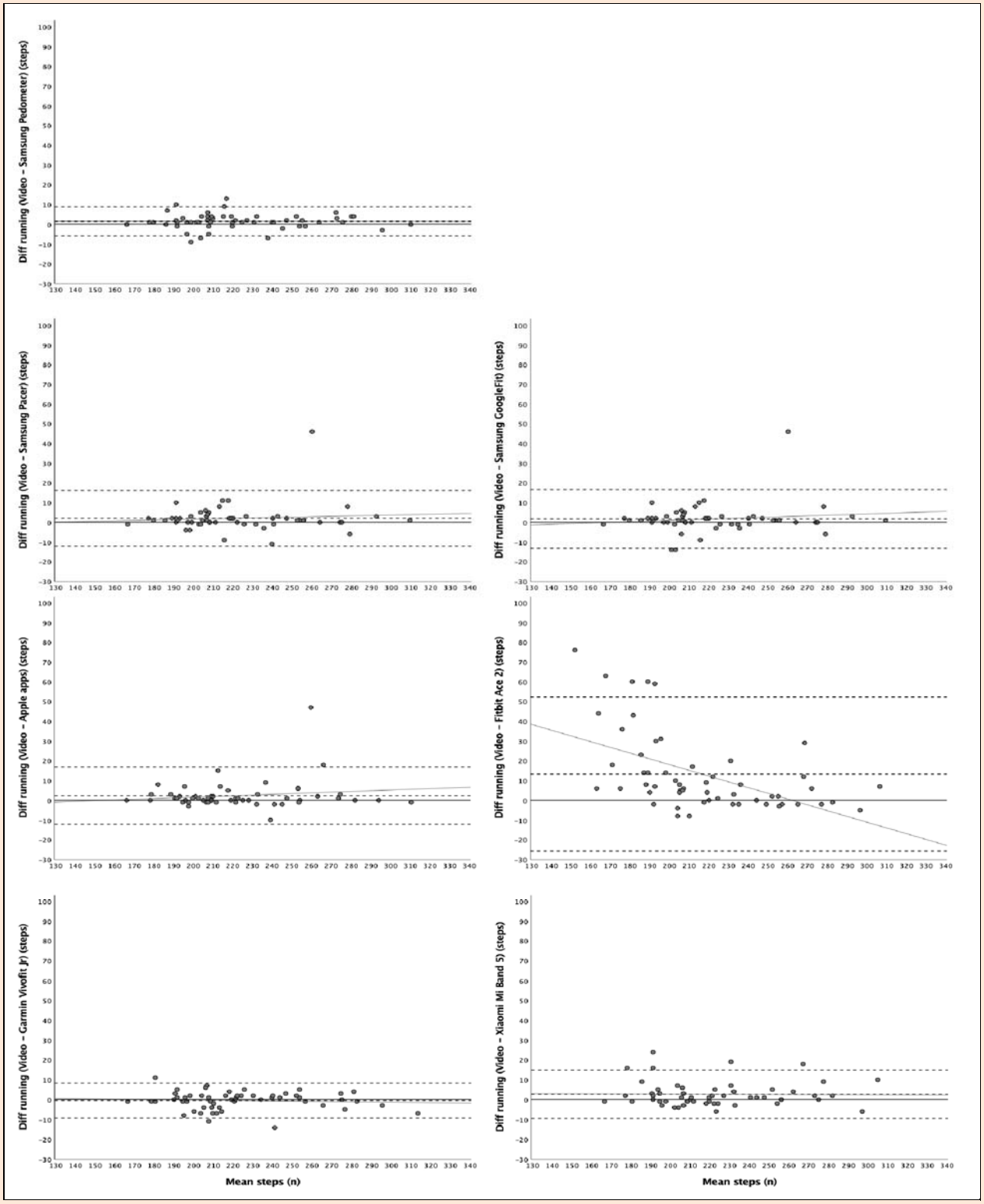

Figure 5. Limits of agreement plots of the consumer-wearable activity trackers for estimating steps during controlled conditions (running condition). The middle-dashed line indicates the mean difference (systematic bias) between step scores assessed by the consumer-wearable activity trackers and the video-based count (gold standard) and the upper and lower dashed lines indicate the limits of agreement (95% confidence interval).

Comparability of the consumer-wearable activity trackers for estimating steps

Table 4 shows the comparability of the consumer-wearable activity trackers for estimating steps during controlled conditions. The results showed that the comparability of the step scores estimated by the activity trackers tended to be higher for slow pace walking, followed by running, normal pace walking and brisk pace walking. Particularly, the results showed that the comparability of the step scores estimated by all the activity trackers in the slow pace walking and running conditions was excellent (e.g., scores inside the 90% CI of the equivalence test, MAPE ≤ 5%, and 95%

CI of the ICC ≥ 0.90), except for the 95% CI of the ICC value with all the comparisons with the Fitbit Ace 2 that was good for slow pace walking and poor for running. The results of the comparability of the step scores estimated by all the activity trackers in the normal pace walking condition was good-excellent, except for the 95% CI of the ICC value with all the comparisons with the Fitbit Ace 2/Xiaomi Mi Band 5 which was poor, as well as between the two wrist-worn activity trackers where it was acceptable. However, while in the brisk pace walking condition the 95% CI of the ICC value with all the comparisons with the three Samsung applications was excellent-good, as well as with the Garmin Vivofit Jr 2 was acceptable, the rest of comparisons were poor (though for all the comparisons the scores were inside the 90% CI of the equivalence test, and the MAPE values were excellent). Pearson's correlation coefficients did not show heteroscedasticity on any walking/running condition (r = - 0.50 - 0.26), except with the comparability between the Garmin Vivofit Jr 2 and Xiaomi Mi Band 5 on the brisk pace walking condition (r = - 0.53), and the Fitbit Ace 2 and Samsung Pedometer/Samsung GoogleFit/Apple applications on the running condition [r = - 0.52 - (- 0.51); Table 3].

Table 4. Comparability of the consumer-wearable activity trackers for estimating steps during controlled conditions (n = 56)

Instrument	Equivalence test (90% CI)	LOA (95% CI)	MAE	MAPE	ICC (95% CI)
Slow pace walking	-50.30, 50.30				
Samsung Pedometer - Samsung Pacer	-1.10, 1.20	0.1 (-9.9, 10.1)	3.6	0.0	0.99 (0.97, 0.99)
Samsung Pedometer - Samsung GoogleFit	-1.07, 1.21	0.1 (-9.9, 10.1)	3.6	0.0	0.99 (0.98, 0.99)
Samsung Pedometer - Apple applications [a]	0.04, 4.25	2.1 (-16.3, 20.5)	4.9	0.0	0.94 (0.91, 0.97)
Samsung Pedometer - Fitbit Ace 2	2.13, 8.13	5.1 (-21.2, 31.4)	9.1	0.0	0.89 (0.80, 0.94)
Samsung Pedometer - Garmin Vivofit Jr 2	-0.57, 1.93	0.7 (-10.3, 11.7)	4.2	0.0	0.98 (0.97, 0.99)
Samsung Pedometer - Xiaomi Mi Band 5	-0.68, 2.21	0.8 (-11.9, 13.5)	4.7	0.0	0.98 (0.96, 0.99)
Samsung Pacer - Samsung GoogleFit	-0.05, 0.09	0.0 (-0.6, 0.6)	0.1	0.0	1.00 (1.00, 1.00)
Samsung Pacer - Apple applications [a]	0.25, 3.93	2.1 (-14.0, 18.2)	4.3	0.0	0.96 (0.93, 0.98)
Samsung Pacer - Fitbit Ace 2	2.23, 7.91	5.1 (-19.8, 30.0)	7.9	0.0	0.90 (0.81, 0.94)
Samsung Pacer - Garmin Vivofit Jr 2	-0.56, 1.81	0.6 (-9.8, 11.0)	4.3	0.0	0.98 (0.97, 0.99)
Samsung Pacer - Xiaomi Mi Band 5	-0.51, 1.94	0.7 (-10.1, 11.5)	4.0	0.0	0.98 (0.97, 0.99)
Samsung GoogleFit - Apple applications [a]	0.24, 3.91	2.1 (-14.0, 18.2)	4.3	0.0	0.96 (0.93, 0.98)
Samsung GoogleFit - Fitbit Ace 2	2.21, 7.90	5.1 (-19.8, 30.0)	7.8	0.0	0.90 (0.81, 0.94)
Samsung GoogleFit - Garmin Vivofit Jr 2	-0.58, 1.79	0.6 (-9.8, 11.0)	4.3	0.0	0.98 (0.97, 0.99)
Samsung GoogleFit - Xiaomi Mi Band 5	-0.53, 1.92	0.7 (-10.1, 11.5)	4.0	0.0	0.98 (0.97, 0.99)
Apple applicationsa - Fitbit Ace 2	0.44, 5.52	3.0 (-19.3, 25.3)	7.6	0.0	0.92 (0.87, 0.95)
Apple applicationsa - Garmin Vivofit Jr	-3.16, 0.24	-1.5 (-16.4, 13.4)	4.1	0.0	0.96 (0.94, 0.98)
Apple applicationsa - Xiaomi Mi Band 5	-3.25, 0.50	-1.4 (-17.9, 15.1)	4.2	0.0	0.96 (0.93, 0.98)
Fitbit Ace 2 - Garmin Vivofit Jr 2	-7.17, -1.72	-4.4 (-28.3, 19.5)	8.3	0.0	0.91 (0.83, 0.95)
Fitbit Ace 2 - Xiaomi Mi Band 5	-6.99, -1.72	-4.4 (-27.5, 18.7)	7.9	0.0	0.92 (0.85, 0.95)
Garmin Vivofit Jr 2 - Xiaomi Mi Band 5	-1.07, 1.25	0.1 (-10.1, 10.3)	4.0	0.0	0.98 (0.97, 0.99)
Normal pace walking	-44.52, 44.52				
Samsung Pedometer - Samsung Pacer	0.46, 3.04	1.8 (-9.6, 13.2)	4.0	0.0	0.97 (0.94, 0.98)
Samsung Pedometer - Samsung GoogleFit	0.48, 3.44	2.0 (-10.9, 14.9)	4.5	0.0	0.96 (0.93, 0.98)
Samsung Pedometer - Apple applications [a]	1.37, 4.78	3.1 (-11.8, 18.0)	5.1	0.0	0.94 (0.89, 0.97)
Samsung Pedometer - Fitbit Ace 2	8.10, 13.83	11.0 (-14.1, 36.1)	12.5	0.0	0.80 (0.37, 0.92)
Samsung Pedometer - Garmin Vivofit Jr 2	-2.05, 1.87	-0.1 (-17.3, 17.1)	6.2	0.0	0.94 (0.89, 0.96)
Samsung Pedometer - Xiaomi Mi Band 5	7.21, 15.22	11.2 (-23.9, 46.3)	12.8	0.0	0.74 (0.45, 0.86)
Samsung Pacer - Samsung GoogleFit	-0.43, 0.85	0.2 (-5.5, 5.9)	0.5	0.0	0.99 (0.99, 1.00)
Samsung Pacer - Apple applications [a]	-0.10, 2.74	1.3 (-11.2, 13.8)	2.6	0.0	0.96 (0.93, 0.98)
Samsung Pacer - Fitbit Ace 2	6.50, 11.92	9.2 (-14.5, 32.9)	10.8	0.0	0.83 (0.50, 0.92)
Samsung Pacer - Garmin Vivofit Jr	-3.54, -0.14	-1.8 (-16.7, 13.1)	4.9	0.0	0.95 (0.91, 0.97)
Samsung Pacer - Xiaomi Mi Band 5	5.58, 13.35	9.5 (-24.6, 43.6)	11.3	0.0	0.76 (0.53, 0.87)
Samsung GoogleFit - Apple applications [a]	-0.51, 2.73	1.1 (-13.0, 15.2)	3.0	0.0	0.95 (0.92, 0.97)
Samsung GoogleFit - Fitbit Ace 2	6.16, 11.84	9.0 (-15.9, 33.9)	11.3	0.0	0.82 (0.53, 0.92)
Samsung GoogleFit - Garmin Vivofit Jr 2	-4.15, 0.05	-2.1 (-20.5, 16.3)	5.2	0.0	0.92 (0.87, 0.95)
Samsung GoogleFit - Xiaomi Mi Band 5	5.28, 13.22	9.3 (-25.6, 44.2)	11.7	0.0	0.75 (0.54, 0.86)
Apple applicationsa - Fitbit Ace 2	4.94, 10.85	7.9 (-18.0, 33.8)	10.7	0.0	0.82 (0.60, 0.91)
Apple applicationsa - Garmin Vivofit Jr 2	-5.08, -1.24	-3.2 (-20.1, 13.7)	5.7	0.0	0.93 (0.87, 0.96)
Apple applicationsa - Xiaomi Mi Band 5	4.00, 12.28	8.1 (-28.2, 44.4)	11.6	0.0	0.74 (0.55, 0.85)
Fitbit Ace 2 - Garmin Vivofit Jr	-13.72, -8.38	-11.1 (-34.4, 12.2)	12.4	0.0	0.82 (0.34, 0.93)
Fitbit Ace 2 - Xiaomi Mi Band 5	-3.19, 3.69	0.3 (-29.9, 30.5)	8.6	0.0	0.86 (0.78, 0.92)
Garmin Vivofit Jr 2 - Xiaomi Mi Band 5	7.72, 14.88	11.3 (-20.1, 42.7)	12.0	0.0	0.78 (0.46, 0.90)

LOA = Limits of Agreement; 90/95% CI = 90/95% Confident Interval; MAE = Mean Absolute Error; MAPE = Mean Absolute Percentage Error; ICC = Intraclass Correlation Coefficient. [a] Apple applications is referred to the three applications activated in the iPhone smartphone (i.e., Pedometer, Pacer, and Apple Health) due to the fact that all of them reported exactly the same steps scores.

Table 4. Continue…

Instrument	Equivalence test (90% CI)	LOA (95% CI)	MAE	MAPE	ICC (95% CI)
Brisk pace walking	-38.97, 38.97				
Samsung Pedometer - Samsung Pacer	-0.09, 2.30	1.1 (-9.5, 11.7)	3.4	0.0	0.95 (0.92, 0.97)
Samsung Pedometer - Samsung GoogleFit	0.33, 3.28	1.8 (-11.1, 14.7)	3.9	0.0	0.93 (0.88, 0.96)
Samsung Pedometer - Apple applications [a]	7.62, 16.13	11.9 (-25.3, 49.1)	13.3	0.1	0.47 (0.16, 0.68)
Samsung Pedometer - Fitbit Ace 2	19.04, 27.67	23.4 (-14.4, 61.2)	24.5	0.1	0.41 (0.00, 0.71)
Samsung Pedometer - Garmin Vivofit Jr 2	-1.11, 3.14	1.0 (-17.6, 19.6)	6.6	0.0	0.86 (0.77, 0.92)
Samsung Pedometer - Xiaomi Mi Band 5	11.49, 22.37	16.9 (-30.7, 64.5)	18.3	0.1	0.39 (0.07, 0.62)
Samsung Pacer - Samsung GoogleFit	-0.31, 1.70	0.7 (-8.1, 9.5)	0.7	0.0	0.97 (0.95, 0.98)
Samsung Pacer - Apple applications [a]	6.47, 15.06	10.8 (-26.8, 48.4)	11.8	0.0	0.48 (0.19, 0.68)
Samsung Pacer - Fitbit Ace 2	17.83, 26.67	22.3 (-16.5, 61.1)	23.6	0.1	0.41 (0.00, 0.70)
Samsung Pacer - Garmin Vivofit Jr	-2.14, 1.96	-0.1 (-18.1, 17.9)	6.2	0.0	0.87 (0.79, 0.92)
Samsung Pacer - Xiaomi Mi Band 5	10.10, 21.55	15.8 (-34.4, 66.0)	18.0	0.1	0.35 (0.07, 0.58)
Samsung GoogleFit - Apple applications [a]	5.59, 14.55	10.1 (-29.1, 49.3)	12.3	0.0	0.47 (0.20, 0.66)
Samsung GoogleFit - Fitbit Ace 2	17.00, 26.10	21.6 (-18.4, 61.6)	23.4	0.1	0.41 (0.00, 0.69)
Samsung GoogleFit - Garmin Vivofit Jr 2	-3.03, 1.46	-0.8 (-20.4, 18.8)	6.8	0.0	0.85 (0.76, 0.91)
Samsung GoogleFit - Xiaomi Mi Band 5	9.24, 21.01	15.1 (-36.4, 66.6)	18.4	0.1	0.35 (0.07, 0.57)
Apple applicationsa - Fitbit Ace 2	6.80, 16.16	11.5 (-29.5, 52.5)	19.3	0.1	0.57 (0.29, 0.74)
Apple applicationsa - Garmin Vivofit Jr 2	-15.17, -6.55	-10.9 (-48.7, 26.9)	14.3	0.1	0.48 (0.19, 0.68)
Apple applicationsa - Xiaomi Mi Band 5	-1.40, 11.51	5.1 (-51.5, 61.7)	18.8	0.1	0.37 (0.12, 0.57)
Fitbit Ace 2 - Garmin Vivofit Jr 2	-26.76, -17.92	-22.3 (-61.1, 16.5)	23.6	0.1	0.41 (0.00, 0.70)
Fitbit Ace 2 - Xiaomi Mi Band 5	-11.16, -1.70	-6.4 (-48.0, 35.2)	15.3	0.1	0.69 (0.52, 0.81)
Garmin Vivofit Jr 2 - Xiaomi Mi Band 5	10.56, 21.26	15.9 (-30.9, 62.7)	20.0	0.1	0.42 (0.10, 0.64)
Running	-33.47, 33.47				
Samsung Pedometer - Samsung Pacer	-1.04, 2.11	0.5 (-13.4, 14.4)	3.7	0.0	0.98 (0.96, 0.99)
Samsung Pedometer - Samsung GoogleFit	-1.41, 1.81	0.2 (-13.9, 14.3)	3.8	0.0	0.97 (0.96, 0.99)
Samsung Pedometer - Apple applications [a]	-0.78, 2.53	0.9 (-13.6, 15.4)	4.2	0.0	0.97 (0.95, 0.98)
Samsung Pedometer - Fitbit Ace 2	7.30, 16.20	11.8 (-27.2, 50.8)	14.2	0.1	0.82 (0.61, 0.91)
Samsung Pedometer - Garmin Vivofit Jr 2	-3.27, -0.73	-2.0 (-13.2, 9.2)	4.5	0.0	0.98 (0.97, 0.99)
Samsung Pedometer - Xiaomi Mi Band 5	-0.33, 2.76	1.2 (-12.3, 14.7)	4.8	0.0	0.98 (0.96, 0.99)
Samsung Pacer - Samsung GoogleFit	-0.88, 0.21	-0.3 (-5.0, 4.4)	0.7	0.0	1.00 (1.00, 1.00)
Samsung Pacer - Apple applications [a]	-0.85, 1.53	0.3 (-10.1, 10.7)	3.7	0.0	0.99 (0.98, 0.99)
Samsung Pacer - Fitbit Ace 2	6.55, 15.88	11.2 (-29.8, 52.2)	14.7	0.1	0.81 (0.62, 0.90)
Samsung Pacer - Garmin Vivofit Jr 2	-4.16, -0.91	-2.5 (-16.8, 11.8)	4.8	0.0	0.97 (0.95, 0.98)
Samsung Pacer - Xiaomi Mi Band 5	-1.56, 2.92	0.7 (-18.9, 20.3)	6.1	0.0	0.95 (0.92, 0.97)
Samsung GoogleFit - Apple applications [a]	-0.59, 1.95	0.7 (-10.5, 11.9)	4.1	0.0	0.98 (0.97, 0.99)
Samsung GoogleFit - Fitbit Ace 2	6.77, 16.34	11.6 (-30.3, 53.5)	15.1	0.1	0.80 (0.60, 0.89)
Samsung GoogleFit - Garmin Vivofit Jr 2	-4.00, -0.39	-2.2 (-18.1, 13.7)	5.3	0.0	0.97 (0.94, 0.98)
Samsung GoogleFit - Xiaomi Mi Band 5	-1.29, 3.32	1.0 (-19.2, 21.2)	6.4	0.0	0.95 (0.91, 0.97)
Apple applicationsa - Fitbit Ace 2	6.39, 15.36	10.9 (-28.5, 50.3)	14.7	0.1	0.82 (0.63, 0.91)
Apple applicationsa - Garmin Vivofit Jr 2	-4.61, -1.14	-2.9 (-18.2, 12.4)	5.2	0.0	0.97 (0.94, 0.98)
Apple applicationsa - Xiaomi Mi Band 5	-1.82, 2.50	0.3 (-18.5, 19.1)	6.1	0.0	0.95 (0.92, 0.97)
Fitbit Ace 2 - Garmin Vivofit Jr 2	-18.15, -9.35	-13.8 (-52.4, 24.8)	14.9	0.1	0.81 (0.52, 0.91)
Fitbit Ace 2 - Xiaomi Mi Band 5	-14.82, -6.26	-10.5 (-47.9, 26.9)	14.0	0.1	0.84 (0.66, 0.92)
Garmin Vivofit Jr 2 - Xiaomi Mi Band 5	1.58, 4.85	3.2 (-11.1, 17.5)	5.6	0.0	0.97 (0.94, 0.98)

LOA = Limits of Agreement; 90/95% CI = 90/95% Confident Interval; MAE = Mean Absolute Error; MAPE = Mean Absolute Percentage Error; ICC = Intraclass Correlation Coefficient. [a] Apple applications is referred to the three applications activated in the iPhone smartphone (i.e., Pedometer, Pacer, and Apple Health) due to the fact that all of them reported exactly the same steps scores

Discussion

Criterion-related validity of the consumer-wearable activity trackers for estimating steps

The findings of the present study showed that the criterion-related validity of the step scores estimated by the three Samsung applications were excellent in all of the four walking/running conditions. Similarly, the criterion-related validity results of the steps estimated by the Garmin Vivofit Jr 2 was excellent, except for the 95% CI of the ICC value on the brisk pace walking condition that was good. However, although for rest of consumer-wearable activity trackers scores were inside the 90% CI of the equivalence test and MAPE values ranged from good to excellent, in the present study some poor ICC outcomes were observed. For instance, while for the Apple applications most of the criterion-related validity results were also excellent, on the brisk pace walking condition it was poor. Similarly, while the criterion-related validity results of the steps estimated by the Xiaomi Mi Band 5 were good-excellent on slow pace walking and running, on the normal and brisk pace walking conditions the ICC outcomes were poor. Finally, the criterion-related validity results of the steps estimated by the Fitbit Ace 2 with the ICC ranged from poor to acceptable.

In spite of the increasing use of smartphone PA applications and wrist-worn activity trackers, today there is still a lack of substantial evidence regarding their criterion-related validity in primary schoolchildren (Fuller et al., 2020; Gorzelitz et al., 2020; Johnston et al., 2021). Prior studies on the criterion-related validity of steps estimated

by wrist-worn activity trackers in primary schoolchildren under controlled conditions showed similar outcomes to the present study. For instance, Godino et al. (2020) studied the criterion-related validity of the Fitbit Charge HR (non-dominant wrist) in primary schoolchildren (mean = 9.9, 9 - 11 years) under controlled and semi free-living conditions (14 structured activities, including sitting, stationary cycling, treadmill walking/running, stair walking, outdoor walking and agility drills), for which they used a person-worn video camera (GoPro Hero) mounted on a harness as the gold standard (two observers). Similar to the results of the present study with the Fitbit Ace 2, Godino et al. (2020) observed that, on average, with the MAPE while the Fitbit Charge HR had a good criterion-related validity for estimating steps under the 14 activities (9.9%), the largest disagreement was found during fast walking/running [LOA = 20.5 (-19,6, 60.6)]. Likewise, these authors also found that the Fitbit Charge HR underestimated step counts [mean of the 14 activities, LOA = 11.8 (8.1, 15.6)].

In the same way, Sun et al. (2022) studied the criterion-related validity of the Fitbit Ace (left wrist) and Moki (right wrist) in primary and secondary schoolchildren (mean = 13.0, 11 - 13 years) under controlled conditions (3 walking activities), for which they used the smartphone camera (iPhone 8) mounted on a tripod as the gold standard (two observers). Similar to the findings of the present study with the Fitbit Ace 2, Sun et al. (2022) also observed that with the MAPE the Fitbit Ace had a good-excellent criterion-related validity for estimating steps (9.5, 3.1 and 5.3%), but it underestimated step counts [LOA = 30.0 (-44.1, 104.1), 3.0 (-21.3, 27.9), and 13.0 (-32.2, 57.3)]. However, these authors found that the Moki had an excellent criterion-related validity for estimating steps (e.g., MAPE = 4.0/3.9/3.0%; systematic bias = 1.0/-4.0/-6.0). On the other hand, to our knowledge, there is no prior study examining the criterion-related validity of Garmin, Xiaomi or any other brand wrist-worn activity trackers for estimating steps in primary schoolchildren under controlled conditions. Likewise, as far as we know, no previous topic-related studies were carried out with smartphone PA applications in this population.

Although validity results depend on the population and testing conditions and, therefore, should not be generalized, due to the limited evidence on the criterion-related validity of wrist-worn activity trackers and smartphone PA applications for estimating steps in primary schoolchildren under controlled conditions, the findings of the present study have also been compared with available literature with young people (under 18 years) under controlled conditions and with primary schoolchildren under free-living conditions. As regards previous studies under controlled conditions, to our knowledge, only Viciana et al. (2022) examined the criterion-related validity of steps estimated by consumer-wearable activity trackers (wrist-worn activity trackers, in non-dominant wrist: Xiaomi Mi Band 5, Samsung Galaxy Watch Active 2, and Apple Watch Series 5; the same PA applications than in the present study) in secondary students (mean = 14.7, 12 - 18 years) under controlled conditions (200-m course at slow, normal and brisk pace walking, and running), for which they used a digital video camera (Go Pro Hero 7) with a tripod situated in the middle of the sports court as the gold standard (two observers). Similar to the results of the present study, Viciana et al. (2022) observed that although for the examined consumer-wearable activity trackers, scores were inside the 90% CI of the equivalence test and the MAPE values were excellent, some ICC outcomes were poor-acceptable. Moreover, similar to the present study, the above-mentioned authors found that while the criterion-related validity results of the steps estimated by the Xiaomi Mi Band 5 were excellent on slow pace walking and running, on the normal and brisk pace walking conditions the ICC outcomes were acceptable and poor, respectively. Similarly, Viciana et al. (2022) also observed that for the three Samsung applications under the four walking/running conditions had good-excellent ICC outcomes (except for the Samsung Pedometer in normal pace walking condition which was poor). Furthermore, while for the Apple applications most of the criterion-related validity results were excellent, on the slow pace walking condition it was acceptable (in the present study on the brisk pace walking condition it was instead poor). Finally, these authors found that the Samsung Galaxy Watch Active 2 and Apple Watch Series 5 had a good-excellent criterion-related validity for estimating steps, except on the brisk pace walking condition which was poor.

Regarding previous studies examining the criterion-related validity of wrist-worn activity trackers for estimating steps in primary schoolchildren under free-living conditions, to our knowledge, only four previous studies were carried out. Similar to the present study, Mayorga-Vega et al. (2023) examined the validity of the wrist-worn activity trackers Fitbit Ace 2, Garmin Vivofit Jr 2, and Xiaomi Mi Band 5 (non-dominant wrist) in primary schoolchildren (mean = 10.4, 9-12 years), for which they used the ActiGraph wGT3X-BT accelerometer as the reference standard (right hip). Similar to the results of the present study, the above-mentioned authors found that while the validity of the primary schoolchildren's daily steps estimated by the Garmin Vivofit Jr 2 and Xiaomi Mi Band 5 were good and acceptable (e.g., scores inside the 90% CI of the equivalence test, MAPE = 9.6/11.3%, and 95% CI of the ICC = 0.87/0.73), respectively, on the contrary, for the Fitbit Ace 2 it was poor (e.g., scores outside the 90% CI of the equivalence test, MAPE = 21.1%, and 95% CI of the ICC = 0.00). Similarly, while Schmidt et al. (2022) observed that the wrist-worn activity tracker Fitbit (Flex 2; non-dominant wrist) had a poor validity (ActiGraph GT9X accelerometer as the reference standard; right hip) for estimating daily steps (e.g., scores were outside the 90% CI of the equivalence test; MAPE = 45.1%) in primary schoolchildren (mean = 8.1, 6-11 years); Yang et al. (2019) found that the wrist-worn activity tracker Xiaomi (model not reported) had an acceptable validity (ActiGraph GT3X-BT accelerometer as the reference standard; right hip) for estimating daily steps (e.g., systematic bias = 633.5) in primary schoolchildren (mean = 13.0, 10 - 17 years). Finally, Sirard et al. (2017) examined the validity of the Movband 2 (dominant wrist) for estimating daily steps in 6-to-12-year-old primary schoolchildren (mean = 8.6 years) using the ActiGraph GT3X+ accelerometer as the reference standard (right hip). These authors found that the Movband 2

considerably overestimated the primary schoolchildren's daily steps (i.e., 2,190.0 steps). As regards the smartphone PA applications, however, to our knowledge, there is no previous study examining their validity for estimating daily steps in primary schoolchildren under free-living conditions.

The above-mentioned previous studies under free-living conditions found that the validity of wrist-worn activity trackers for estimating steps tended to be lower than under controlled conditions. However, these apparent inconsistences between the findings of the present study (i.e., controlled conditions) and those in free-living conditions are plausible. While in the studies carried out in controlled conditions such as the present study, primary schoolchildren are constrained to predefined activities with stable gait patterns, previous studies under free-living conditions were carried out under a greater variability of motor patterns including a wide range of children's daily life behaviors. Consequently, it is to be expected that the mean error is lower in the first above mentioned case compared with the error in measurement in the second case (Johnston et al., 2021). In this line, systematic reviews have shown that consumer-wearable activity trackers tend to have a higher validity for estimating steps under controlled conditions than under free-living conditions (Fuller et al., 2020; Gorzelitz et al., 2020). Furthermore, criterion-related validity of step counts estimated by consumer-wearable activity trackers should be analyzed by examining the agreement between their scores and those from the "gold standard" (Johnston et al., 2021), that is, an error-free reference standard (Bossuyt et al., 2015). Video-based step counting with at least two observers is widely considered the gold standard (Johnston et al., 2021). However, all the above-mentioned previous studies under free-living conditions were carried out with ActiGraph accelerometers as the reference standard, that is, a non-error free method (normally underestimating step counts, especially in slow pace walking) for assessing step counts among primary schoolchildren (Rosenkranz et al., 2010).

The findings of the present study indicate that the criterion-related validity of step scores estimated by activity trackers tended to be higher during slow pace walking and running conditions compared to normal and brisk pace walking in primary schoolchildren. Several factors could contribute to the observed differences in criterion-related validity across various walking and running conditions, such as algorithms or biomechanics of movement. These factors collectively underestimate the need for nuanced algorithm design and consideration of biomechanical variations across different walking and running conditions to enhance the overall validity of activity trackers in step counting in primary schoolchildren.

As it was mentioned before, the findings of the present study showed that the criterion-related validity of the step scores estimated by the three Samsung applications and the Garmin Vivofit Jr 2 were good-excellent on all four walking/running conditions. Given that most primary schoolchildren now own smartphones that they carry with them throughout the day (Spanish National Institute of Statistics, 2023) and the three studied Samsung applications are freely available, these PA applications would hold a significant advantage as they do not require to purchase any specific device for monitoring and promoting PA. In both the present study and the previous topic-related study with secondary students (Viciana et al., 2022) the criterion-related validity of PA applications was examined with the smartphone placed in simulated front trouser pockets. However, in real life primary schoolchildren also placed their smartphones in many other places such as back trouser pockets, schoolbags, or in their hands. Likewise, in many moments smartphones are placed somewhere away from the body. Therefore, the criterion-related validity of the step scores estimated by the three studied Samsung applications in real life could be considerably lower. In contrast, wrist-worn activity trackers such as the Garmin Vivofit Jr 2 offer a significant advantage due to their high wear compliance (Fairclough et al., 2016). Moreover, these consumer-wearable activity trackers are always worn in the same location, which aligns with their placement on the non-dominant wrist as it was examined in the present study. Thus, in real life the validity of the Garmin Vivofit Jr 2 potentially would be considerably higher than of the Samsung applications.

Comparability of the consumer-wearable activity trackers for estimating steps

The findings of the present study showed that the comparability of the step scores estimated between the three Samsung applications were excellent on all four walking/running conditions (except between the Samsung Pedometer and GoogleFit on the brisk pace walking condition that with the ICC was good). Additionally, the comparability of the step scores estimated by the three Samsung applications and Garmin Vivofit Jr 2 was good-excellent, except on the brisk pace walking condition that with the ICC was acceptable. However, although for the rest of the consumer-wearable activity trackers scores were inside the 90% CI of the equivalence test and MAPE values were excellent, in the present study some poor ICC outcomes were observed. Particularly, the comparability with the ICC was poor for all the comparisons with the Apple applications under the brisk pace walking condition; with the Xiaomi Mi Band 5 under the normal (except with the Xiaomi Mi Band 5 that was just acceptable) and brisk pace walking conditions; and with the Fitbit Ace 2 under the normal and brisk pace walking, and running conditions.

Although the use of different consumer-wearable activity trackers to monitor and promote primary schoolchildren's steps is commonly used in contexts with economic constraints such as in Physical Education where primary schoolchildren use their own devices (Viciana et al., 2022), to our knowledge, unfortunately, there are no previous studies about the comparability of wrist-worn activity trackers and smartphone PA applications in primary schoolchildren under controlled conditions. As far as we know, only Mayorga-Vega et al. (2023) examined the comparability of steps estimated by the wrist-worn activity trackers Fitbit Ace 2, Garmin Vivofit Jr 2, and Xiaomi Mi Band 5 (non-dominant wrist) in primary schoolchildren, but under free-living conditions. The above-mentioned authors found that while the comparability of the daily step scores estimated by the Garmin Vivofit Jr 2 and Xiaomi Mi

Band 5 were acceptable-excellent (e.g., scores inside the 90% CI of the equivalence test, MAPE ≤ 5%, and 95% CI of the ICC ≥ 0.70), it was poor for those two with the Fitbit Ace 2 (e.g., 95% CI of the ICC < 0.70). Similarly, in the present study apart from when scores were inside the 90% CI of the equivalence test and MAPE values were excellent under all the conditions, the comparability with the ICC was excellent between the Garmin Vivofit Jr 2 and Xiaomi Mi Band 5 under the slow pace walking and running conditions, but it was poor under the normal and brisk pace walking conditions. Moreover, for all the conditions except the slow pace walking the comparability with the ICC was poor for those two with the Fitbit Ace 2 (exceptionally was just acceptable with the Xiaomi Mi Band 5 under the normal pace walking conditions).

Viciana et al. (2022) also examined the comparability of wrist-worn activity trackers (Samsung Galaxy Watch Active 2, Apple Watch Series 5, and Xiaomi Mi Band 5; non-dominant wrist), but in a sample of secondary students and under free-living conditions. Although the Xiaomi Mi Band 5 and Samsung Galaxy Watch Active 2 had a good-excellent comparability [e.g., MAPE = 8.4; ICC = 0.98 (0.91 - 0.99)], the comparability between those with the Apple Watch Series 5 was poor (e.g., MAPE = 19.4/23.3%). However, as far as we know, no previous studies about the comparability of step scores estimated by smartphone PA applications was carried out in this population.

Considering that the three Samsung applications and the Garmin Vivofit Jr 2 were comparable for estimating daily steps, apart from the price, technical characteristics, and options offered by the different devices, this could also be an important reason to select one or another for a particular aim (Viciana et al., 2022). For instance, battery duration, attractive screen, goal settings, reminders, or the data registered in the application, among others, could be essential to consider (Casado-Robles et al., 2022). However, as it was mentioned before, due to the fact that in real life the criterion-related validity of the Garmin Vivofit Jr 2 potentially would be considerably higher than that of the three studied Samsung applications, essentially those applications could be not comparable with the Garmin Vivofit Jr 2. On the other hand, as regards the wrist-worn activity trackers, in settings such as in Physical Education where the only economical possible way is that students use their own device (i.e., already purchased), for instance, the findings of the present study showed that the studied wrist-worn activity trackers could not be used interchangeably to monitor and promote daily steps among primary schoolchildren.

Strengths and limitations

An important strength of the present study was being, to our knowledge, the first one to examine the criterion-related validity of the steps estimated in primary schoolchildren under controlled conditions by recent models of wrist-worn activity trackers specially designed for children (i.e., Fitbit Ace 2 and Garmin Vivofit Jr 2). Moreover, as far as we know, it is also the first study to examine the criterion-related validity of the steps estimated by PA smartphone applications in primary schoolchildren under controlled conditions. Furthermore, to the best of our knowledge, the present study is the first one to examine the comparability of step counts estimated by smartphone PA applications and wrist-worn activity trackers among primary schoolchildren. This is another important issue for feasibility reasons because in contexts such as Physical Education or large-scale research studies, it is common to use different consumer-wearable activity trackers for different participants (Brodie et al., 2018; Creaser et al., 2022). Finally, an important consideration during the validation of consumer-wearable activity trackers is the validity of the criterion measure to which they are being compared (Johnston et al., 2021). If the criterion measure is not sufficiently valid, then criterion standard bias may be present (Umemneku Chikere et al., 2019), limiting the value that may be gained from such a study. Thus, due to the fact that the video-based step counting with at least two observers is considered the gold standard (i.e., an error-free reference standard) (Johnston et al., 2021), the criterion measure used in the present study represents another strength. Therefore, the present study allows for addressing important gaps in the scientific literature to date.

However, the present study has some limitations. Firstly, a non-probability and relatively small sample has been used, which limits the generalizability of the obtained outcomes to the particular studied setting (i.e., primary schoolchildren with similar characteristics). However, due to the human, time and material resource restrictions, a probability and larger sample could not be examined. Secondly, the best-practice protocol for the validation of steps estimated by consumer-wearable activity trackers should be conducted under controlled, semi-free living, and free-living conditions (Johnston et al., 2021). Although the controlled testing condition represents the first stage in the multistage protocols for the best-practice validation of steps estimated by consumer-wearable activity trackers (Johnston et al., 2021), focusing only on controlled conditions may fail in the ecological validation of these devices under free-living conditions (Johnston et al., 2021). However, due to the human, time and material resource restrictions, the criterion-related validity under semi-free living, and especially free-living conditions could not be examined. Although previous studies have shown that it is feasible using a body worn camera as a gold standard under free-living conditions (Kelly et al., 2015), from the researcher's perspective, feasibility is compromised by the processing time that video recording requires (e.g., 600 - 900 minutes of video to be analyzed by two independent observers for each participant and day) (Johnston et al., 2021).

Consequently, future studies should examine the criterion-related validity (i.e., with a gold standard) of steps scores in primary schoolchildren under semi-free living, and free-living conditions. Moreover, since other PA outputs such as heart rate, distance, or energy expenditure, are commonly used in the consumer-wearable activity trackers, future studies should also examine the criterion-related validity (i.e., with a gold standard) of those PA outputs in primary schoolchildren.

Conclusion

The PA applications Pedometer, Pedometer Pacer Health, and Google Fit for Android (using the Samsung Galaxy S20+ placed in the trousers pockets) and the wrist-worn activity tracker Garmin Vivofit Jr 2 (on the non-dominant wrist) showed good-excellent criterion-related validity for estimating steps in primary schoolchildren under controlled conditions. Moreover, those devices were comparable for estimating steps in primary schoolchildren under controlled conditions. However, the Apple applications and the wrist-worn activity trackers Fitbit Ace 2 and Xiaomi Mi Band 5 have not shown an acceptable criterion-related validity for estimating steps, as well as they not being comparable, on some walking/running conditions. However, due to the fact that in real life primary schoolchildren also place their smartphones in other places (e.g., schoolbags or hands), including somewhere away from the body, the criterion-related validity of the Samsung applications could be considerably lower. In contrast, because in real life, wrist-worn activity trackers are usually worn on the wrist the same as it was examined in the present study, in real life the validity of the Garmin Vivofit Jr 2 potentially would be considerably higher than that of the Samsung applications. The findings of the present study highlight the potential of the Garmin Vivofit Jr 2 for monitoring primary schoolchildren's steps under controlled conditions. Future studies should examine the criterion-related validity of steps scores estimated by PA applications and wrist-worn activity trackers in primary schoolchildren under free-living conditions.

Acknowledgements

Authors thank Aliisa Hatten for the English revision. This study is part of I + D + I Fit-Person project (Reference: PID2019110179GB-100), funded by the MCIN/AEI/10.13039/501100011033. There is no conflict of interest. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The present study complies with the current laws of the country in which it was performed. The datasets generated and analyzed during the current study are not publicly available but are available from the corresponding author, who was an organizer of the study.

References

Althof, T., Sosič, R., Hicks, J., King, A., Delp, S. and Leskovec, J. (2017) Large-scale physical activity data reveal worldwide activity inequality. *Nature* **547**, 336-339. https://doi.org/10.1038/nature23018

Arifin, W. (2018) A web-based sample size calculator for reliability studies. *Education in Medicine Journal* **10**, 67-76. https://doi.org/10.21315/eimj2018.10.3.8

Atkinson, G. and Nevill, A. (1998) Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine* **26**, 217-238. https://doi.org/10.2165/00007256-199826040-00002

Benítez-Porres, J., Alvero-Cruz, J., Sardinha, L., López-Fernández, I. and Carnero, E. (2016) Cut-off values for classifying active children and adolescents using the Physical Activity Questionnaire: PAQ-C and PAQ-A. *Nutricion Hospitalaria* **33**, 1036-1044. https://doi.org/10.20960/nh.564

Bland, J. and Altman, D. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* **1**, 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bland, J. M. and Altman, D. G. (1995) Comparing methods of measurement: why plotting difference against standard method is misleading. *The Lancet* **346**, 1085-1087. https://doi.org/10.1016/S0140-6736(95)91748-9

Bossuyt, P., Reitsma, J., Bruns, D., Gatsonis, C., Glasziou, P., Irwig, L., Lijmer, J., Moher, D., Rennie, D., de Vet, H., Kressel, H., Rifai, N., Golub, R., Altman, D., Hooft, L., Korevaar, D. and Cohen, J. (2015) STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *British Medical Journal* **351**, 5527. https://doi.org/10.1136/bmj.h5527

Brodie, M. A., Pliner, E. M., Ho, A., Li, K., Chen, Z., Gandevi, S. C. and Lord, S. R. (2018) Big data vs accurate data in health research: Large-scale physical activity monitoring, smartphones, wearable devices and risk of unconscious bias. *Medical Hypotheses* **119**, 32-36. https://doi.org/10.1016/j.mehy.2018.07.015

Casado-Robles, C., Viciana, J., Guijarro-Romero, S. and Mayorga-Vega, D. (2022) Effects of consumer-wearable activity tracker-based programs on objectively measured daily physical activity and sedentary behavior among school-aged children: A systematic review and meta-analysis. *Sports Medicine - Open* **8**, 18. https://doi.org/10.1186/s40798-021-00407-6

Cohen, J. (1992) A power primer. *Quantitative Methods in Psychology* **112**, 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Cole, T., Bellizzi, M., Flegal, K. and Dietz, W. (2000) Establishing a standard definition for child overweight and obesity worldwide: International survey. *British Medical Journal BMJ* **320**, 1240-1243. https://doi.org/10.1136/bmj.320.7244.1240

Colley, R., Janssen, I. and Tremblay, M. (2012) Daily step target to measure adherence to physical activity guidelines in children. *Medicine and Science in Sports and Exercise* **44**, 977-982. https://doi.org/10.1249/MSS.0b013e31823f23b1

Craig, C., Cameron, C., Griffiths, J. and Tudor-Locke, C. (2010) Descriptive epidemiology of youth pedometer-determined physical activity: CANPLAY. *Medicine and Science in Sports and Exercise* **42**, 1639-1643. https://doi.org/10.1249/MSS.0b013e3181d58a92

Creaser, A. v, Frazer, M. T., Costa, S., Bingham, D. D. and Clemes, S. A. (2022) The use of wearable activity trackers in schools to promote child and adolescent physical activity: A descriptive content analysis of school staff's perspectives. *International Journal of Environmental Research and Public Health* **19**, 14067. https://doi.org/10.3390/ijerph192114067

Crossley, S., McNarry, M., Rosenberg, M., Knowles, Z., Eslambolchilar, P. and Mackintosh, K. (2019) Understanding youths' ability to interpret 3D-printed physical activity data and identify associated intensity levels: Mixed-methods study. *Journal of Medical Internet Research* **21**, e11253. https://doi.org/10.2196/11253

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

Fairclough, S., Noonan, R., Rowlands, A., Van Hees, V., Knowles, Z. and Boddy, L. (2016) Wear compliance and activity in children wearing wrist- and hip-mounted accelerometers. *Medicine and Science in Sports and Exercise* **48**, 245-253. https://doi.org/10.1249/MSS.0000000000000771

Fuller, D., Colwell, E., Low, J., Orychock, K., Tobin, M. A., Simango, B., Buote, R., van Heerden, D., Luan, H., Cullen, K., Slade, L. and Taylor, N. G. A. (2020) Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: Systematic review. *JMIR Mhealth Uhealth* **8**, e18694. https://doi.org/10.2196/18694

Gil-Espinosa, F. J., Nielsen-Rodríguez, A., Romance, R. and Burgueño, R. (2022) Smartphone applications for physical activity promotion from physical education. *Education and Information Technologies* **27**, 11759-11779. https://doi.org/10.1007/s10639-022-11108-2

Giurgiu, M., Kolb, S., Nigg, C., Burchartz, A., Timm, I., Becker, M., Rulf, E., Doster, A.-K., Koch, E., Bussmann, J. B. J., Nigg, C., Ebner-Priemer, U. W. and Woll, A. (2022) Assessment of 24-hour physical behaviour in children and adolescents via wearables: a systematic review of free-living validation studies. *BMJ Open Sport* **8**, e001267. https://doi.org/10.1136/bmjsem-2021-001267

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

Gorzelitz, J., Farber, C., Gangnon, R. and Cadmus-Bertram, L. (2020)

Accuracy of wearable trackers for measuring moderate- to vigorous-intensity physical activity: A systematic review and meta-analysis. *Journal for the Measurement of Physical Behaviour* **3**, 346-357. https://doi.org/10.1123/jmpb.2019-0072

Hartung, V., Sarshar, M., Karle, V., Shammas, L., Rashid, A., Roullier, P., Eilers, C., Mäurer, M., Flachenecker, P., Pfeifer, K. and Tallner, A. (2020) Validity of consumer activity monitors and an algorithm using smartphone data for measuring steps during different activity types. *International Journal of Environmental Research and Public Health* **17**, 9314. https://doi.org/10.3390/ijerph17249314

Henriksen, A., Mikalsen, M., Woldaregay, A., Muzny, M., Hartvigsen, G., Hopstock, L. and Grimsgaard, S. (2018) Using fitness trackers and smartwatches to measure physical activity in research: Analysis of consumer wrist-worn wearables. *Journal of Medical Internet Research* **20**, e110. https://doi.org/10.2196/jmir.9157

Johnston, W., Judice, P. B., Molina García, P., Mühlen, J. M., Lykke Skovgaard, E., Stang, J., Schumann, M., Cheng, S., Bloch, W., Brønd, J. C., Ekelund, U., Grøntved, A., Caulfield, B., Ortega, F. B. and Sardinha, L. B. (2021) Recommendations for determining the validity of consumer wearable and smartphone step count: expert statement and checklist of the INTERLIVE network. *British Journal of Sports Medicine* **55**, 780-793. https://doi.org/10.1136/bjsports-2020-103147

Kelly, P., Thomas, E., Doherty, A., Harms, T., Burke, Ó., Gershuny, J. and Foster, C. (2015) Developing a method to test the validity of 24 hour time use diaries using wearable cameras: A feasibility pilot. *Plos One* **10**, e0142198. https://doi.org/10.1371/journal.pone.0142198

Koo, T. and Li, M. (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* **15**, 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B., Hróbjartsson, A., Roberts, C., Shoukri, M. and Streiner, D. (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies* **48**, 661-671. https://doi.org/10.1016/j.ijnurstu.2011.01.017

Laricchia, F. (2023a) Global smartphone sales to end users 2007-2022. Statista. Available from URL: https://www.statista.com/statistics/263437/global-smartphone-sales-to-end-users-since-2007/ [Accessed 15 October 2023].

Laricchia, F. (2023b) Wearables - statistics & facts. Statista. Available from URL: https://www.statista.com/topics/1556/wearable-technology/#topicOverview [Accessed 15 October 2023].

Maher, C., Ryan, J., Ambrosi, C. and Edney, S. (2017) Users' experiences of wearable activity trackers: a cross-sectional study. *BMC Public Health* **17**, 880. https://doi.org/10.1186/s12889-017-4888-1

Mayorga-Vega, D., Casado-Robles, C., Guijarro-Romero, S. and Viciana, J. (2023) Validity of activity wristbands for estimating daily physical activity in primary schoolchildren under free-living conditions: School-Fit study. *Frontiers in Public Health* **11**, 1211237. https://doi.org/10.3389/fpubh.2023.1211237

Mayorga-Vega, D., Casado-Robles, C., López-Fernández, I. and Viciana, J. (2021) A comparison of the utility of different step-indices to translate the physical activity recommendation in adolescents. *Journal of Sports Sciences* **39**, 469-479. https://doi.org/10.1080/02640414.2020.1826667

Mayorga-Vega, D., Casado-Robles, C., López-Fernández, I. and Viciana, J. (2022) Activity wristband-based physical activity recommendations in young people. *Science & Sports* **37**, 303-315. https://doi.org/10.1016/j.scispo.2021.05.007

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M. and de Vet, H. C. W. (2010) The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research* **19**, 539-549. https://doi.org/10.1007/s11136-010-9606-8

Nunnally, J. (1978) Psychometric theory. 2nd edition. New York: McGraw-Hill.

Oliveira, L. C., Ferrari, G. L. de M., Araújo, T. L. and Matsudo, V. (2017) Overweight, obesity, steps, and moderate to vigorous physical activity in children. *Revista de Saúde Pública*, **51**, 38. https://doi.org/10.1590/s1518-8787.2017051006771

Poitras, V., Gray, C., Borghese, M., Carson, V., Chaput, J., Janssen, I., Katzmarzyk, P., Pate, R., Connor Gorber, S., Kho, M., Sampson, M. and Tremblay, M. (2016) Systematic review of the relationships between objectively measured physical activity and health indicators in school-aged children and youth. *Applied Physiology, Nutrition, and Metabolism* **41**, 197-239. https://doi.org/10.1139/apnm-2015-0663

Rosenkranz, R., Rosenkranz, S. and Weber, C. (2010) Validity of the Actigraph accelerometer step count function in children. *Journal of Science and Medicine in Sport* **13**, 97-98. https://doi.org/10.1016/j.jsams.2010.10.667

Schmidt, M. D., Rathbun, S. L., Chu, Z., Boudreaux, B. D., Hahn, L., Novotny, E., Johnsen, K. and Ahn, S. J. (2023) Agreement between Fitbit and ActiGraph estimates of physical activity in young children. *Measurement in Physical Education and Exercise Science* **27**, 171-180. https://doi.org/10.1080/1091367X.2022.2125319

Sirard, J. R., Masteller, B., Freedson, P. S., Mendoza, A. and Hickey, A. (2017) Youth oriented activity trackers: Comprehensive laboratory- and field-based validation. *Journal of Medical Internet Research* **19**, 250. https://doi.org/10.2196/jmir.6360

Spanish National Institute of Statistics. (2023) Use of ICT products by children 6 to 15 years old. Spanish National Institute of Statistics. Available from URL: https://www.ine.es/jaxi/Datos.htm?tpx=50168 [Accessed 15 October 2023] (In Spanish)

Stewart, A., Marfell-Jones, M., Olds, T. and De Ridder, J. (2011) International standards for anthropometric assessment. New Zealand: International Society for the Advancement of Kinanthropometry.

Sun, X., Adams, S. A., Li, C., Booth, J. N., Robertson, J. and Fawkner, S. (2022) Validity of the Fitbit Ace and Moki devices for assessing steps during different walking conditions in young adolescents. *Pediatric Exercise Science* **34**, 1-5. https://doi.org/10.1123/pes.2021-0026

Umemneku Chikere, C. M., Wilson, K., Graziadio, S., Vale, L. and Allen, A. J. (2019) Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard - An update. *Plos One* **14**, e0223832. https://doi.org/10.1371/journal.pone.0223832

Viciana, J., Casado-Robles, C., Guijarro-Romero, S. and Mayorga-Vega, D. (2022) Are wrist-worn activity trackers and mobile applications valid for assessing physical activity in high school students? Wearfit study. *Journal of Sports Science and Medicine* **21**, 356-375. https://doi.org/10.52082/jssm.2022.356

Willmott, C. and Matsuura, K. (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* **30**, 79-82. https://doi.org/10.3354/cr030079

World Health Organization. (2020) WHO guidelines on physical activity and sedentary behaviour. Geneva: World Health Organization.

Yang, X., Jago, R., Zhang, Q., Wang, Y., Zhang, J. and Zhao, W. (2019) Validity and reliability of the wristband activity monitor in free-living children aged 10−17 years. *Biomedical and Environmental Sciences* **32**, 812-822.

Key points

- The PA applications Pedometer, Pedometer Pacer Health, and Google Fit for Android and the wrist-worn activity tracker Garmin Vivofit Jr 2 showed good-excellent criterion-related validity for estimating steps in primary schoolchildren under controlled conditions.
- The PA applications Pedometer, Pedometer Pacer Health, and Google Fit for Android and the wrist-worn activity tracker Garmin Vivofit Jr 2 were comparable for estimating steps in primary schoolchildren under controlled conditions.
- The Apple applications and the wrist-worn activity trackers Fitbit Ace 2 and Xiaomi Mi Band 5 have not shown an acceptable criterion-related validity for estimating steps, as well as not being comparable, on some walking/running conditions.
- Since in real life primary schoolchildren also place their smartphones in other places (e.g., schoolbags, hands or even somewhere away from the body), the criterion-related validity of the Garmin Vivofit Jr 2 potentially would be considerably higher than that of the Samsung applications.

AUTHOR BIOGRAPHY

Daniel MAYORGA-VEGA
Employment
Departamento de Didáctica de las Lenguas, las Artes y el Deporte, Facultad de Ciencias de la Educación, Universidad de Málaga, Málaga, España.
Degree
PhD
Research interests
Measurement and evaluation, health-enhancing physical activity, sedentary behavior, health-related physical fitness, Physical Education-based interventions, motivation toward physical activity.
E-mail: dmayorgavega@uma.es

Carolina CASADO-ROBLES
Employment
Department of Physical Education and Sport, University of Granada, Spain.
Degree
PhD
Research interests
Measurement and evaluation, health-enhancing physical activity, sedentary behavior, Physical Education-based interventions, motivation toward physical activity.
E-mail: ccasado@ugr.es

Santiago GUIJARRO-ROMERO
Employment
Department of Didactic of Musical, Plastic and Corporal Expression, University of Valladolid, Valladolid, Spain
Degree
PhD
Research interests
Measurement and evaluation, health-enhancing physical activity, sedentary behavior, Physical Education-based interventions, motivation toward physical activity.
E-mail: santigr93@gmail.com

Jesús VICIANA
Employment
Department of Physical Education and Sport, University of Granada, Spain.
Degree
PhD
Research interests
Measurement and evaluation, health-enhancing physical activity, sedentary behavior, Physical Education-based interventions, motivation toward physical activity.
E-mail: jviciana@ugr.es

✉ **Santiago Guijarro-Romero**
Department of Didactic of Musical, Plastic and Corporal Expression, Faculty of Education and Social Work, University of Valladolid, Paseo de Belén, 1, 47011, Valladolid, Spain