Validity and Reliability of Wearable Devices during Self-Paced Walking, Jogging and Overground Skipping

James W. Navalta 1, Dustin W. Davis 1, Bryson Carrier 1, Elias M. Malek 1, Nicole Vargas 1, Jorge Perdomo Rodriguez 1, Bianca Weyers 1, Katherine Carlos 1, Myranda Peck 1

1University of Nevada, Las Vegas, Department of Kinesiology and Nutrition Sciences, Las Vegas, Nevada, USA

Abstract

Wearable technology can track unusual exercise, providing data for improving fitness. The aim of the study was to determine validity and reliability during walking, jogging, and skipping. Eighteen volunteers completed 5 min self-paced activities interspersed with 5 min rest. Variables and devices were step count (Garmin Instinct), estimated energy expenditure (Garmin Instinct, Polar Vantage M2), and heart rate (Garmin Instinct, Polar Vantage M2, Polar OH1, Polar Verity Sense). Validity measures were mean absolute percent error (MAPE) and Lin's Concordance (CCC), and reliability were coefficient of variation (CV), and intraclass correlation (ICC). Thresholds were MAPE ≤5%, CCC≥0.90, CV≤10%, ICC≥0.70. Garmin Instinct step count during skipping was not considered valid (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Instinct (MAPE=90.2%, CCC=0.008) or reliable (CV=6%, ICC CI=0.4). Energy expenditure during skipping was not valid or reliable in the Garmin Inst.\n
Keywords: wearable activity tracker, unusual exercise, accuracy, consistency

Introduction

Wearable technology, particularly activity trackers, have sprung up in popularity in recent years. These wearable devices are designed to monitor and track physical activity, including steps taken, estimated energy expenditure, and heart rate (Bunn, Navalta, Fountaine, & Reece, 2018). Systematic reviews consistently report that using wearable activity trackers leads to improvements in weight loss, waist circumference, and body mass index (Yen & Chiu, 2019). Wearing an activity tracker and receiving daily feedback on progress can lead to increased physical activity and reduced sedentary behavior (Jakicic et al., 2016). However, the boundless enthusiasm for wearable devices should be tempered, as the benefits of using activity trackers may depend on motivation and level of engagement with the device (Cadmus-Bertram, Marcus, Patterson, Parker, & Morey, 2015).

Wearable technology has been increasingly used to monitor and track physical activity beyond traditional exercise activities, such as in water-based sports, dancing, and video gaming. Activity trackers have been proposed to obtain metrics during water-based activities such as kayaking and canoeing (Umek & Kos, 2018; Liu, Wang, Qiu, Zhang, & Hao, 2018). However, the boundless enthusiasm for wearable devices should be tempered, as the benefits of using activity trackers may depend on motivation and level of engagement with the device (Cadmus-Bertram, Marcus, Patterson, Parker, & Morey, 2015).
2021). Similarly, wearable sensors can be used to assess activity during both social and jazz dancing activities (Stancin & Tomazić, 2022), making it a useful tool for dancers who must perform repetitive leaping motions to track their progress and improve their performance. Furthermore, virtual reality gaming has gained popularity as an alternative form of exercise, and wearable technology such as virtual reality headsets and motion capture devices can be used to monitor and track physiological responses during these activities (Cao, Xie, & Chen, 2019). Overall, wearable technology has shown potential in monitoring and tracking physical activity in a variety of unusual exercise activities, providing useful data for individuals to optimize their performance and improve their overall fitness levels. However, further research is needed to explore the accuracy and reliability of these devices in diverse settings and unusual activities.

Bounding straight to the issue, overground skipping is a non-traditional exercise that has not been investigated to date in the context of wearable technology devices. Most humans transition from walking to running as a faster method of travel and exercise (Navalta, Davis, Carrier, Sertic, & Cater, 2021), however, running requires greater energy expenditure than walking (Harrell et al., 2005). Some people, particularly children, perform skipping as a transitory movement as an alternative to walking and running (Navalta et al., 2021). The available literature centers on rope skipping, rather than overground skipping (Verdel et al., 2022; Yongmao & Yuxin, 2023). One study evaluated skipping in 20-sec bouts as obtained from a novel smart patch (Verdel et al., 2022), and a recent investigation was conducted on rope skipping and deep learning algorithms (Yongmao & Yuxin, 2023). However, overground skipping is different in that the action involves forward propulsion that is absent in rope skipping, which may be responsible for half of the energy cost of walking (Gottschall & Kram, 2005). Thus, the evaluation of wearable devices to return valid and reliable measurements during overground skipping should be conducted.

The purpose of this investigation was to determine the validity and reliability of wearable technology devices during an overground skipping activity. While many studies are designed to determine validity, there is a need to also obtain reliability metrics (Carrier, Barrios, Jolley, & Navalta, 2020). Previous research has shown wearable devices to have varied levels of validity, with heart rate generally being acceptable, energy expenditure estimation unacceptable, and step count falling in between (Bunn et al., 2018). Based on previous literature and testing performed in our laboratory, it was hypothesized that during overground skipping, step count would be reliable but not valid, estimates of energy expenditure would be neither valid nor reliable, and heart rate would be both valid and reliable. These findings are important, as this is the first study to report validity and reliability of wearable devices during an overground skipping activity.

### Methods

#### Participants

Eighteen participants (female n=10, male n=8, transgender, intersex, or other n=0) provided written informed consent that was approved by the Institutional Review Board (IRB approval# UNLV-2022-80) and volunteered for the study. A power analysis was performed using pilot data, indicating the need for at least eleven participants (coefficient of determination $r^2=0.57$, correlation $r$ effect size=0.755, $a=0.05$, $b=0.80$). Participants were screened and deemed not to require medical clearance to complete exercise according to the American College of Sports Medicine preparticipation health screening recommendations. The mean demographic information included: age=26±8 years, body mass=73±11.9 kg, height=66.4±3.5 cm, and Body Mass Index=25.6±3.0 kg·m$^{-2}$. Participants self-identified ethnic group included African American (n=1), White (n=10), Hispanic (n=3), Polynesian (n=1), and Southeast Asian (n=3).

#### Exercise Bouts

After confirming connection to the criterion and experimental devices, participants performed a self-paced walk back and forth through an indoor hallway along a 61-meter track, followed by 5 minutes (min) of seated rest. All devices were reset during the period of rest (i.e., new activity recording session with the same demographic and anthropometric data in the watch; not a factory reset). Participants then completed 5 min of self-paced running back and forth through the same indoor hallway, followed by 5 min of seated rest. All devices were again reset during the period of rest. Finally, participants performed 5 min of self-paced overground skipping back and forth through the same indoor hallway. Step count was collected via manual clicker step counters by two independent observers during the overground skipping bout only. The step count was the arithmetic mean of the observer's manual counts. A step was defined as the completion of a unilateral stride before transferring motion to the contralateral
The protocol is similar in terms of timing and administration to other investigations conducted by the laboratory group (Montes & Navalta, 2019; Navalta et al., 2019; Montes, Tandy, Young, Lee, & Navalta, 2022).

**Devices**

Polar H10: Although the use case specific to overground skipping has not been determined, the Polar H10 chest strap has been shown in other settings to have acceptable reliability (Speer, Semple, Naumovski, & McKune, 2020) and to be valid compared to electrocardiography (Gilgen-Ammann, Schweizer, & Wyss, 2019). The Polar H10 is an electrocardiogram-based heart rate sensor secured around the chest at the level of the xyphoid process. The H10 contains plastic electrodes on the underside of the strap that detects heart rate. The sensor materials include acrylonitrile butadiene styrene (ABS), ABS plus glass fiber (ABS+GF), polycarbonate, and stainless steel, and the strap material is composed of 38% polyamide, 29% polyurethane, 20% elastane, 13% polyester, and silicone prints. The H10 has a sampling frequency of 1000 Hertz (Hz).

Polar OH1: The Polar OH1 is a photoplethysmography (PPG) device that uses an optical heart rate sensor worn on the upper arm. The OH1 is an optical heart rate sensor worn on the upper arm. The sensor materials include ABS, ABS+GF, poly(methyl methacrylate) (PMMA), and steel use stainless (SUS) 316. The device was positioned so that the sensor was on the underside of the armband and firmly against the skin. The OH1 has a sample rate of 135 Hz.

Polar Verity Sense: The Polar Verity Sense is also a PPG device worn on the upper arm. The sensor materials include ABS, ABS+GF, PMMA, and SUS 316. The device was positioned with the sensor on the underside of the armband and firmly against the skin. The Verity Sense has a sample rate of 135 Hz.

Garmin Instinct: The Garmin Instinct is a PPG device. The physical size is 46 x 46 x 12.5 millimeters (mm) with a mass of 45 grams (g). The optical heart rate sensor is worn on the upper arm. The sensor materials include acrylonitrile butadiene styrene (ABS), ABS plus glass fiber (ABS+GF), polycarbonate, and stainless steel, and the strap material is composed of 38% polyamide, 29% polyurethane, 20% elastane, 13% polyester, and silicone prints. The Garmin Instinct has a sampling frequency of 1000 Hertz (Hz).

Polar Vantage M2: The Polar Vantage M2 is a PPG device. The physical size is 45 x 45 x 15.3 mm with a mass of 52 g. The optical heart rate sensor is worn on the wrist and employs Precision PrimeTM heart rate sensor technology. The Vantage M2 is reported to have a sample rate of 135 Hz.

**Statistical Analysis**

Reported measures associated with validity include mean absolute percent error (MAPE), Lin’s Concordance Correlation Coefficient (CCC), and the mean absolute error (MAE). The equations for these metrics were input into an Excel spreadsheet (Microsoft Excel for Mac version 16.66.1, Redmond, WA). For validity thresholds our laboratory group has used a MAPE value ≤ 5%, and a CCC ≥ 0.90 (J. W. Navalta et al., 2020). A device meeting both thresholds was considered evidence in support of validity. The Bland-Altman analysis was used to determine agreement. Bias and limits of agreement were determined using the blandr analysis in jamovi (version 2.3.19.0).

Reported measures associated with reliability include the coefficient of variation (CV), and two-way mixed model with absolute agreement intraclass correlation coefficient (ICC). The CV was determined using Excel (Microsoft Excel for Mac version 16.66.1, Redmond, WA), and the ICC (single measures) using SPSS Statistics (IBM SPSS Statistics, version 28.0.1.0, Chicago, IL). For non-laboratory settings a threshold of ≤ 10% for CV, and ≥ 0.70 for ICC, with a lower bound of the 95% CI ≥ 0.70 has been used (Navalta et al., 2019). A device meeting both thresholds was considered evidence in support of reliability.

**Results**

**Step Count**

Step count was obtained only from the Garmin Instinct during the skipping trial because the other devices did not return the measure. Manual step count was not obtained from one participant and was not recorded from the Garmin Instinct for two other participants. Thus, validity measures reflect data points from 30 distinct bouts, and reliability measures from 16 pairs (see table 1). Based on the predefined definition of a step, the device did not meet any of the threshold measures for validity and underestimated the step count by nearly double (see Table 1). The Garmin Instinct met the reliability thresholds for CV and ICC but was not considered reliable because the lower end of the 95% CI did not meet the established threshold.

<table>
<thead>
<tr>
<th>Skipping</th>
<th>Validity Data Points</th>
<th>Average MAPE (%)</th>
<th>Bias Limits of Agreement</th>
<th>CCC 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual Count</td>
<td>30</td>
<td>507.3 (32.9)</td>
<td>90.2</td>
<td>-499.8</td>
</tr>
<tr>
<td>Garmin Instinct</td>
<td>30</td>
<td>957.2 (139.9)</td>
<td>60.4</td>
<td>-350.9</td>
</tr>
<tr>
<td>Reliability Data Points</td>
<td>Device 1</td>
<td>Device 2</td>
<td>CV (%)</td>
<td>ICC 95% CI</td>
</tr>
<tr>
<td>Garmin Instinct</td>
<td>16</td>
<td>933.7 (165.3)</td>
<td>1000.9 (103.4)</td>
<td>6.0</td>
</tr>
</tbody>
</table>

Average is the arithmetic mean (standard deviation) measurement of step count. MAPE = mean absolute percent error, CCC = Lin’s Concordance Correlation Coefficient, CV = coefficient of variation, ICC = Intraclass correlation coefficient, CI = Confidence interval of the ICC. Values noted in bold and italics meet the predetermined threshold for validity or reliability.

**Energy Expenditure**

There was no missing energy expenditure data from the criterion device or any of the wearable activity trackers. Validity and reliability measures for wrist-worn devices (Garmin Instinct, Polar Vantage M2) during walking, jogging, and skipping are provided in Table 2. The biceps-worn devices did not return an estimate of energy expenditure. Neither the Garmin Instinct nor the Polar Vantage M2 met the predetermined validity thresholds when returning estimates of energy expenditure across any bout of exercise. The Polar Vantage M2 met the reliability thresholds during self-paced walking and jogging, but not overground skipping.
Heart rate measures were not obtained from the Polar OH1 for one participant during the walking trial. There was no missing heart rate data from the criterion device or any of the other wearable activity trackers. All devices met the pre-established validity thresholds during the self-paced walking trial (see Table 3). One wrist-worn device (Garmin Instinct) and one biceps-worn device (Polar Verity Sense) met the thresholds for reliability during walking (see Table 3) and can be considered both valid and reliable during this type of exercise.

Table 3. Heart rate during self-paced walking, jogging, and skipping bouts. Criterion (H10), and validity and reliability measures of wrist-worn (Garmin Instinct, Polar Vantage M2) and biceps-worn (Polar Verity Sense, Polar OH1) devices.

<table>
<thead>
<tr>
<th>Walking</th>
<th>Validity</th>
<th>Data Points</th>
<th>Average MAPE (%)</th>
<th>Bias</th>
<th>Limits of Agreement</th>
<th>CCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSMED K5</td>
<td>36</td>
<td>104.5 (17.1)</td>
<td>2.8</td>
<td>−1.0</td>
<td>7.0 to −9.1</td>
<td>0.97</td>
</tr>
<tr>
<td>Garmin Instinct</td>
<td>36</td>
<td>105.5 (16.6)</td>
<td>3.3</td>
<td>−0.2</td>
<td>13.0 to −13.4</td>
<td>0.91</td>
</tr>
<tr>
<td>Polar Vantage M2</td>
<td>36</td>
<td>104.7 (14.2)</td>
<td>0.7</td>
<td>0.3</td>
<td>2.2 to −1.7</td>
<td>0.99</td>
</tr>
<tr>
<td>Polar Verity Sense</td>
<td>36</td>
<td>104.3 (16.7)</td>
<td>2.1</td>
<td>0.5</td>
<td>14.9 to −13.9</td>
<td>0.91</td>
</tr>
<tr>
<td>Polar OH1</td>
<td>35</td>
<td>104.5 (18.0)</td>
<td>2.1</td>
<td>0.5</td>
<td>14.9 to −13.9</td>
<td>0.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Jogging</th>
<th>Validity</th>
<th>Data Points</th>
<th>Average MAPE (%)</th>
<th>Bias</th>
<th>Limits of Agreement</th>
<th>CCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSMED K5</td>
<td>36</td>
<td>107.3 (16.5)</td>
<td>106.6 (15.9)</td>
<td>2.2</td>
<td>0.96</td>
<td>0.91 to 0.99</td>
</tr>
<tr>
<td>Garmin Instinct</td>
<td>36</td>
<td>105.0 (11.1)</td>
<td>106.0 (16.4)</td>
<td>3.3</td>
<td>0.78</td>
<td>0.50 to 0.91</td>
</tr>
<tr>
<td>Polar Vantage M2</td>
<td>18</td>
<td>105.9 (16.3)</td>
<td>105.6 (16.0)</td>
<td>0.3</td>
<td>0.99</td>
<td>0.99 to 1.0</td>
</tr>
<tr>
<td>Polar OH1</td>
<td>17</td>
<td>106.3 (18.7)</td>
<td>107.0 (16.2)</td>
<td>2.7</td>
<td>0.842</td>
<td>0.62 to 0.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Skipping</th>
<th>Validity</th>
<th>Data Points</th>
<th>Average MAPE (%)</th>
<th>Bias</th>
<th>Limits of Agreement</th>
<th>CCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSMED K5</td>
<td>36</td>
<td>154.9 (13.9)</td>
<td>147.2 (15.8)</td>
<td>5.9</td>
<td>32.8 to −17.5</td>
<td>0.55</td>
</tr>
<tr>
<td>Garmin Instinct</td>
<td>36</td>
<td>148.2 (15.1)</td>
<td>151.7 (13.7)</td>
<td>2.0</td>
<td>10.3 to −4.0</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Average is the arithmetic mean (standard deviation) measurement of kilocalories expended. MAPE = mean absolute percent error, CCC = Lin’s Concordance Correlation Coefficient, CV = coefficient of variation, ICC = Intraclass correlation coefficient, CI = Confidence interval of the ICC. Values noted in bold and italics meet the predetermined threshold for validity or reliability.
Discussion

The purpose of this investigation was to determine the validity and reliability of wearable technology devices to return step count, energy expenditure, and heart rate during walking, jogging, and overground skipping. It was hypothesized that, during overground skipping, step count would be reliable but not valid, energy expenditure would be neither valid nor reliable, and heart rate would be both valid and reliable. First, the hypothesis for step count was partially supported because step count during self-paced skipping was neither valid nor reliable. Second, the hypothesis for energy expenditure was supported because estimated energy expenditure during self-paced skipping was neither valid nor reliable. Third, the hypothesis for heart rate was supported for biceps-worn devices (valid and reliable) but not wrist-worn technology (neither valid nor reliable).

The conclusions drawn from step count obtained during skipping are obtained from a single wrist-worn wearable device and should be considered with caution. The device did not meet the validity thresholds, likely because of how a step was defined in the current investigation (as a complete unilateral stride before transferring motion to the contralateral side of the body). The definition did not account for the foot tap down during a stride that many participants performed during the skipping motion, and likely accounted for the very large bias that was observed. It may be useful for future researchers interested in step count during overground skipping to account for the foot tap, as it appears at least one wearable device registers this action. It is proposed that users of consumer activity trackers may utilize this knowledge to their advantage, to increase step count during wellness program challenges by performing overground skipping as their preferred method of transportation, which would conceivably double the step count compared to the more mundane method of walking. The practical implications of this finding should not be skipped over.

Heart rate was observed to be both valid and reliable in all devices during the walking bout. When wrist-worn devices (Garmin Instinct, Polar Vantage M2) are considered, the finding did not extend to activities of greater relative intensity, being considered neither valid nor reliable during jogging and overground skipping. This finding aligns with a study using a different wrist-based device, the Garmin fēnix 5, which reported moderate validity measures during sitting and walking, but poor validity measures during increased intensities of exercise (Duking et al., 2020). However, an earlier Polar wrist-worn model, the Polar Vantage V, was reported to have acceptable MAPE values for heart rate during running and cycling at low and high self-selected intensities (Haji-Boutros, Landry-Duval, Comtois, Gouspillou, & Karelis, 2023). The specific wrist-based devices used in the current investigation have little published literature outside of conference abstracts, many of which are from our laboratory group. The same holds true for one of the biceps-worn devices, the Polar Verity Sense, in that published validation literature currently exists only in abstract form (Bodell et al., 2021). The other biceps-worn device, the Polar OH1, is reported to have acceptable validity during treadmill (MAPE between 0.2 and 1.9%) and cycle exercise (MAPE between 0.6 and 3.9%) (Muggeridge et al., 2021), spin bike activities (mean bias less than 1 bpm) (Hettiarachchi, Hanoun, Nahavandi, & Nahavandi, 2019),
front crawl swimming at various intensities (ICC between 0.72 and 0.96) (Oldstad & Zinner, 2020), and endurance sports (difference from criterion < 5%), as well as acceptable reliability during endurance sports (ICC = 0.99) (Hermand, Cassirame, Ennequin, & Hue, 2019). Based on the results of the current investigation, biceps-worn devices may be preferable for individuals who are interested in obtaining both valid and reliable heart rate measures in non-traditional activities such as overground skipping.

It has been noted that estimated energy expenditure in wearable devices during exercise validation studies is generally poor (Bunn et al., 2018). As described with heart rate above, the specific devices used in the current investigation have not been reported in the published literature, and the same holds true for energy expenditure. The existing literature details opportunities for improvement when energy expenditure is considered. Physical activity energy expenditure in wearable devices including the Garmin vivofit, Jawbone Up24, and Fitbit Flex, compared to doubly labeled water were deemed unacceptable (MAPE range for 12 devices 19.4% to 100.2%) (Murakami et al., 2019). Estimated energy expenditure returned from the Garmin vivo HR+ and Fitbit Charge 2 were evaluated against COSMED units (K4, K5) during progressive exercise tests to volitional fatigue (treadmill, cycle ergometer) and MAPE did not meet validity threshold (18.9% to 43.5%) (Reddy et al., 2018). During trail running, the Hexoskin biometric shirt did not return correlated energy expenditure measures compared to the COSMED K4 (r = 0.058) (Tanner et al., 2016). The Apple watch 6 (14.9% to 47.8% MAPE), Polar Vantage V (15.6% to 34.6% MAPE), and Fitbit Sense (17.8% to 45.1% MAPE) all displayed unacceptable MAPE during walking and jogging, and ending with overground skipping. Because the intent of the study was to obtain concurrent validity and reliability measures, it is believed that this approach is acceptable. However, it is unknown whether scheduling the overground skipping trial last affected mechanics of the skill, and whether potential fatigue may have induced extraneous movements that decreased validity and or reliability measurements in certain devices. Future studies employing various exercise modalities should not skip over randomizing the order of exercise bouts.

In conclusion, the modality of overground skipping is an area of research that represents boundless possibilities. In the context of the current investigation, several findings are reported on specific wearable devices for the first time. Regarding heart rate, the Garmin Instinct, Polar Vantage M2, Polar Verity Sense, and Polar OH1 met all thresholds to be considered both valid and reliable during self-paced walking. Only the biceps-worn devices (Polar OH1, Polar Verity Sense) met the heart rate thresholds for self-paced jogging and overground skipping. The Garmin Instinct was the only device to return step count, and it was neither valid nor reliable during overground skipping according to our predetermined definition of a step. While the Polar Vantage M2 was reliable for energy expenditure during walking and jogging, wrist-worn devices (Garmin Instinct, Polar Vantage M2) were neither valid nor reliable in returning estimated energy expenditure during overground skipping. From a wider perspective, wearable device algorithms for estimating energy expenditure should continue to be refined until they return the same level of accuracy as what is currently observed for heart rate, and to a lesser extent step count. Overground skipping may be an excellent unusual exercise for the continued testing of these measures as leaps and bounds are made in the development wearable activity devices.


