Assessment of Peak Oxygen Uptake with a Smartwatch and its Usefulness for Training of Runners

Authors
Peter Düking[1], Bas Van Hooren[2], Billy Sperlich[1]

Affiliations
1 Integrative and Experimental Exercise Science, Department of Sport Science, University of Würzburg, Würzburg, Germany
2 Department of Nutrition and Movement Sciences, NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University Medical Centre+, Maastricht, Netherlands

Key words
data-guided training, digital health, digital training, eHealth, innovation, technology, wearable, mHealth

accepted 20.10.2021
published online 30.01.2022

Bibliography
Int J Sports Med 2022; 43: 642–647
DOI 10.1055/a-1686-9068
ISSN 0172-4622

Georg Thieme Verlag KG, Rüdigerstraße 14,
70469 Stuttgart, Germany

Correspondence
Dr. Peter Düking
Integrative and Experimental Exercise Science,
Department of Sport Science
University of Würzburg
Judenbühlweg 11
97072 Würzburg
Germany
Tel.: +49/931/31 84792
peter.dueking@uni-wuerzburg.de

ABSTRACT

Peak oxygen uptake ($\dot{V}O_{2peak}$) is an important factor contributing to running performance. Wearable technology may allow the assessment of $\dot{V}O_{2peak}$ more frequently and on a larger scale. We aim to i) validate the $\dot{V}O_{2peak}$ assessed by a smartwatch (Garmin Forerunner 245), and ii) discuss how this parameter may assist to evaluate and guide training procedures. A total of 23 runners (12 female, 11 male; $\dot{V}O_{2peak}$: 48.6 ± 6.8 ml·min^{-1}·kg^{-1}) visited the laboratory twice to determine their $\dot{V}O_{2peak}$ during a treadmill ramp test. Between laboratory visits, participants wore a smartwatch and performed three outdoor runs to obtain $\dot{V}O_{2peak}$ values provided by the smartwatch. The $\dot{V}O_{2peak}$ obtained by the criterion measure ranged from 38 to 61 ml·min^{-1}·kg^{-1}. The mean absolute percentage error (MAPE) between the smartwatch and the criterion $\dot{V}O_{2peak}$ was 5.7 %. The criterion measure revealed a coefficient of variation of 4.0% over the VO2peak range from 38–61 ml·min^{-1}·kg^{-1}. MAPE between the smartwatch and criterion measure was 7.1, 4.1 and −6.2 % when analyzing $\dot{V}O_{2peak}$ ranging from 39–45 ml·min^{-1}·kg^{-1}, 45–55 ml·min^{-1}·kg^{-1} or 55–61 ml·min^{-1}·kg^{-1}, respectively.

Introduction

Peak oxygen uptake ($\dot{V}O_{2peak}$) is extensively investigated among individuals of different age, gender, and performance levels [1–4] and is a key component of endurance performance in heterogeneous populations. Although $\dot{V}O_{2peak}$ does not predict performance in homogeneous groups of athletes (i. e., elite level) and while changes in $\dot{V}O_{2peak}$ allows predicting some but not all changes in endurance performance [5], an exceptionally high $\dot{V}O_{2peak}$ constitutes a prerequisite for competitive success in endurance athletes [3, 6]. Based on the peak values, percentages of $\dot{V}O_{2peak}$ are often applied in sports practice to prescribe training intensity, although they are subject to current scientific debate [7].

Maintaining or improving $\dot{V}O_{2peak}$ is an important goal in the training process of runners. Since individuals show considerable inter- and intra-individual physiological responses to the same training procedures [2, 8], frequent evaluation of the effectiveness of training procedures and responsive adjustments of training procedures are required by evaluating important performance indicators (such as $\dot{V}O_{2peak}$ and others).

The accurate assessment of $\dot{V}O_{2peak}$ requires i) time-consuming and expensive laboratory setup for gas exchange measurement, ii) specialized laboratory staff, and iii) an all-out effort by the participant. These disadvantages impair frequent assessment of $\dot{V}O_{2peak}$, especially for recreational runners without access to such equipment. These limitations might be surpassed by advancements in the field of wearable sensors (e. g., smartwatches) and accompanying machine learning algorithms intended to assess $\dot{V}O_{2peak}$. Wearable sensors used in research settings (e. g., a combination of an accelerometer worn on the tibia and a heart rate sensor) employing a mixed-effects unpenalized linear regression model allow the estimation of $\dot{V}O_{2peak}$ with an error of 4.92 % in the laboratory [9]. Nevertheless, these sensors and algorithms may not be available to the public, and few studies have evaluated the validity of $\dot{V}O_{2peak}$ measurements with end consumer wearables (e. g., smartwatches) [10, 11]. However, frequent hard- and software developments of end consumer devices likely affect data quality, and therefore it is important to regularly evaluate these devices for daily application [12, 13]. Regarding the daily use of this technology and data, another challenge is to interpret and draw physiologically meaningful conclusions for training procedures. In this regard, recreational runners will need some level of knowledge on how to interpret changes in $\dot{V}O_{2peak}$ to guide their training [14].

The goal of the present investigation is twofold: i) to validate the $\dot{V}O_{2peak}$ provided by an end consumer smartwatch (Garmin Forerunner 245) against a common criterion measure, and ii) to briefly discuss the usefulness and shortcomings of $\dot{V}O_{2peak}$ measurements to guide a runner's training.

Materials and Methods

Participants

Twenty-three non-competitive recreational runners (11 men, 12 women, mean age 23 ± 3 years, body height 173 ± 8 cm, body mass 70.1 ± 11.2 kg; $\dot{V}O_{2peak}$: 48.6 ± 6.8 ml/min/kg; training characteristics: 2–3 times per week for 45 min at a self-perceived low intensity) of Caucasian origin were informed about all experimental procedures and provided written consent to participate. The study was approved by the institute's ethical committee and performed in accordance with the declaration of Helsinki and the study follows ethical standards in sport and exercise science research [15].

Experimental procedures

The experimental procedure is illustrated in ▶ **Fig. 1**.

All participants reported twice (7–10 days apart) to the laboratory for assessment of anthropometric data, maximal heart rate, and $\dot{V}O_{2peak}$. Even with gold-standard criterion measures, there is an error stemming from technical error and random within-subject variation [16]. To assess the error of the criterion measure in our sample, we tested each participant twice with the criterion measure in the laboratory. This repeated measure allows calculating i) the mean $\dot{V}O_{2peak}$ values of both laboratory visits, which delivers a better estimation of an individual's $\dot{V}O_{2peak}$; and ii) the reliability of the gold-standard criterion-measures allowing comparison to the validity error between the criterion and the smartwatche-derived $\dot{V}O_{2peak}$.

To assess $\dot{V}O_{2peak}$ provided by the smartwatch, the manufacturer's instructions for use indicate a person should run outdoors for at least 10 min with a heart rate "several minutes" above 70 % of the maximal heart rate [17]. The manufacturer indicates that the $\dot{V}O_{2peak}$ assessment might improve following "a couple" of runs [17]. Therefore, between both laboratory visits, all runners performed three outdoor runs (longer than 30 min) on flat terrain.

Ramp test protocol for assessment of peak oxygen uptake

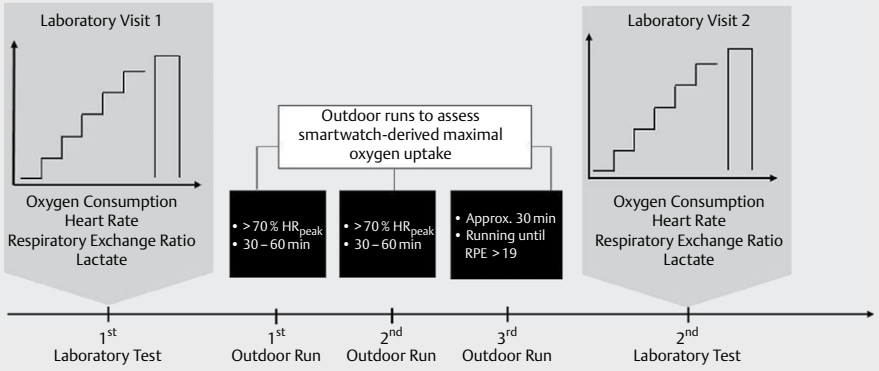
Each participant performed a ramp protocol on a motorized treadmill (Mercury, h/p/cosmos sports and Medical GmbH, Nussdorf-Traunstein, Germany) to assess $\dot{V}O_{2peak}$. Initially the treadmill speed was set to 7 $km \cdot h^{-1}$ increasing every minute by 1 $km \cdot h^{-1}$ until volitional exhaustion. In our experience, this ramp slope (i. e., $km \cdot h^{-1}$ increment) allows recreational runners to reach exhaustion within approximately 10–15 min, which is important for accurate assessment of $\dot{V}O_{2peak}$[7]. Exhaustion was verified if three of the four following criteria were met: 1) plateau in $\dot{V}O_2$, that is, an increase < 1.0 $mL \cdot min^{-1} \cdot kg^{-1}$ despite an increase in velocity; 2) respiratory exchange ratio > 1.1; 3) rating of perceived exertion > 18; and 4) peak blood lactate (peak lactate) > 6 $mmol \cdot L^{-1}$ 30 s after ramp testing. After completion of the ramp test, the participants performed passive recovery for 5 min followed by an instantaneous step increase in running velocity (verification phase) corresponding to 105 % of the velocity achieved during the ramp test. The verification phase ended with each runner's individual volitional exhaustion [18]. The $\dot{V}O_{2peak}$ values, assessed by averaging the last 30 s of the ramp and verification run, were compared [18] and the higher value was used for further analysis.

Assessment of smartwatch derived peak oxygen uptake

Each runner wore two smartwatches, one the left wrist and one on the right. This allowed us to obtain estimates for $\dot{V}O_{2peak}$ from two independent smartwatches at the same time. All participants were instructed to perform three outdoor runs at a constant pace without stopping. To align with the manufacturers recommendations and to ensure that each participant ran "several minutes" above 70 % of peak heart rate (for the first two runs), they all were instructed to run for 30–60 min until exhaustion (i. e., > 18 on the Borg scale). For the third run, the runners were instructed to run for 30 min until fully exerted. We assessed the level of exhaustion by the rating of perceived exertion (RPE) [19], which all runners reported approximately 20 min after completing the running session.

Criterion measure

A portable breath-by-breath analyzer (Metamax 3B, CORTEX Biophysik GmbH, Leipzig, Germany) served as the criterion measure. The oxy-



▶ **Fig. 1** Experimental procedure. Laboratory visit: Ramp protocol to assess maximal oxygen uptake by the criterion. Initial speed set to 7 km·h^{-1}, increasing every minute by 1 km·h^{-1}. Outdoor runs to assess smartwatch derived maximal uptake.

gen sensor of this portable breath-by-breath gas analyzer provides reliable data with technical measurement error below 2 % [20].

Smartwatch

An end consumer smartwatch (Forerunner 245, Garmin, Olathe, USA) employing an optical heart rate sensor as well as a GPS receiver unit was used for this study. We chose the optical heart rate sensor (and not an electrical chest belt sensor) as the optical sensors are becoming more readily available and when optical sensors prove scientific trustworthiness, it is likely that runners will choose this type of sensor due to greater comfort compared to a chest strap. The smartwatch was programmed as indicated by the manufacturer. We did not enter the participants' maximum heart rate into the software since many recreational runners do not know their actual individual maximum heart rate. The exact algorithms of $\dot{V}O_{2peak}$ assessment are not disclosed by the manufacturer, yet it is indicated that reliable heart and GPS-derived velocity data segments from individual runs are used to estimate $\dot{V}O_{2peak}$ [21].

Statistical analysis

A dependent *t*-test (performed in the Statistica Software package for Windows Version 7.1) assessed the difference in peak oxygen uptake between the two exercise tests. An alpha level of ≤ 0.05 was considered statistically significant.

Reliability of the criterion measure

As previously performed [22], reliability of the criterion measure $\dot{V}O_{2peak}$ was calculated as the percentage change in the mean (CM %) and typical error (TE %) expressed as a coefficient of variation (CV %), calculated as SD of the percentage change scores between repeated measures divided by the square root of 2. The intraclass correlation coefficient (ICC, 3.1) was calculated and interpreted according to [23] in order to examine overall group-level association. ICC values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively [23]. For all measures, the corresponding 95 % CI were calculated.

Validity analysis comparing the end consumer smartwatch against the criterion measure

To investigate the validity of the $\dot{V}O_{2peak}$ provided by the smartwatch, we averaged the $\dot{V}O_{2peak}$ of the three runs. We also split the sample in runners with low ($\dot{V}O_{2peak}$ ≤ 45 ml·kg^{-1}·min^{-1}), medium ($\dot{V}O_{2peak}$ 45–55 45 ml·kg^{-1}·min^{-1}), and high ($\dot{V}O_{2peak}$ ≥ 55 45 ml·kg^{-1}·min^{-1}) $\dot{V}O_{2peak}$ to evaluate whether the validity differed between the subgroups. As no international standards exist for thresholds of low, medium, and high $\dot{V}O_{2peak}$ categories these levels are arbitrary. To additionally examine the validity of several runs, we calculated all statistical parameters mentioned in this section for $\dot{V}O_{2peak}$ values that were given for each of the three outdoor runs.

As previously performed, mean absolute percent errors (MAPE) were calculated to provide an indicator of overall measurement error [24]. MAPE was calculated as the average of absolute difference between the smartwatch and the criterion measure divided by the criterion measure value, multiplied by 100.

Bland–Altman plots display the corresponding 95 % limits of agreement and fitted lines (from regression analyses between mean and difference) with their corresponding parameters (i. e., intercept and slope). A fitted line that provides a slope of 0 and an intercept of 0 exemplifies perfect agreement [24].

Results

All descriptive statistics of the laboratory tests and the outdoor runs are summarized in ▶ **Tables 1** and **2**.

The mean $\dot{V}O_{2peak}$ of the two criterion tests were 47.5 ± 6.8 ml·min^{-1}·kg^{-1} and 49.7 ± 7.2 ml·min^{-1}·kg^{-1} (average of laboratory assessed $\dot{V}O_{2peak}$: 48.6 ± 7.0 ml·min^{-1}·kg^{-1}) and this difference was significant (p = 0.0003) between the two tests.

We had to discard three $\dot{V}O_{2peak}$ estimations derived from the smartwatches due to handling errors occurring with the smartwatch. The mean $\dot{V}O_{2peak}$ estimated by the smartwatch after the first, second, and third run as well as the mean of all runs was 49.1 ± 4.6 ml·min^{-1}·kg^{-1}, 49.0 ± 4.7 ml·min^{-1}·kg^{-1}, 48.3 ± 9.0 ml·min^{-1}·kg^{-1}, 49.1 ± 4.6 ml·min^{-1}·kg^{-1}, respectively.

Overall, the criterion measure showed a CM % of 4.6 (95 %CI − 3.1 to 7.4), a TE % as CV % of 4.0 (95 %CI − 0.7 to 4.7). The TE % as CV % of 4.0 corresponds to an error of 1.96 ml·min^{-1}·kg^{-1}. The ICC of 0.943 (95 %CI 0.736 to 0.982) indicates excellent reliability. When splitting the $\dot{V}O_{2peak}$ into subgroups of lower ($\dot{V}O_{2peak}$ ≤ 45 ml·min^{-1}·kg^{-1}; n = 12), medium ($\dot{V}O_{2peak}$ 45–55 ml·min^{-1}·kg^{-1}; n = 13), and higher ($\dot{V}O_{2peak}$ ≥ 55 ml·min^{-1}·kg^{-1}; n = 13) $\dot{V}O_{2peak}$, the criterion measure showed a TE % as CV % of 2.6 % (95 %CI − 0.1 to 2.7); 3.5 (95 %CI 0.0 to 3.8) and 4.0 % (95 %CI 0.8 to 6.3).

When averaging the $\dot{V}O_{2peak}$ values of all three runs, the smartwatch showed a MAPE of 5.7 % (corresponding to an error of 2.80 ml·min^{-1}·kg^{-1}). When the $\dot{V}O_{2peak}$ provided by the smartwatch following the first, second, and third outdoor run were compared, the MAPE was 5.7 % (corresponding to an error of 2.80 ml·min^{-1}·kg^{-1}), 5.6 % (corresponding to an error of 2.70 ml·min^{-1}·kg^{-1}) and 5.6 % (corresponding to an error of 2.70 ml·min^{-1}·kg^{-1}).

The Bland-Altman plot is displayed in ▶ **Fig. 2**.

When the $\dot{V}O_{2peak}$ were split into subgroups of lower ($\dot{V}O_{2peak}$ ≤ 45 ml·min^{-1}·kg^{-1}), medium ($\dot{V}O_{2peak}$ 45–55 ml·min^{-1}·kg^{-1}), and higher ($\dot{V}O_{2peak}$ ≥ 55 ml·min^{-1}·kg^{-1}) $\dot{V}O_{2peak}$, the smartwatch showed a MAPE of 7.1 % (corresponds to an error of 3.48 ml·min^{-1}·kg^{-1}), 4.1 % (corresponds to an error of 2.01 ml·min^{-1}·kg^{-1}) and − 6.2 % (corresponds to an error of − 3.04 ml·min^{-1}·kg^{-1}), respectively.

Discussion

The primary goal of the present investigation was to validate the $\dot{V}O_{2peak}$ provided by an end consumer smartwatch (Garmin Forerunner 245) against a common criterion measure. The two main findings are:

1) Over the $\dot{V}O_{2peak}$ range of 38 to 61 ml·min^{-1}·kg^{-1}(as measured by the criterion measure), the overall MAPE between the smartwatch and the criterion is 5.7 % (~2.8 ml·min^{-1}·kg^{-1}). The MAPE does not seem to decrease when performing one, two or three runs.
2) When clustering the runners' $\dot{V}O_{2peak}$ (i. e., 39 to 45 ml·min^{-1}·kg^{-1}, 45 to 55 ml·min^{-1}·kg^{-1}, and 55 to 61 ml·min^{-1}·kg^{-1}) the MAPE is 7.1 % (~ 3.5 ml·min^{-1}·kg^{-1}), 4.1 % (~ 2.0 ml·min^{-1}·kg^{-1}) and − 6.2 % (~ − 3.0 ml·min^{-1}·kg^{-1}), indicating that within the lower $\dot{V}O_{2peak}$ category, the smartwatch tends to overestimate the runners' actual $\dot{V}O_{2peak}$, whereas within the higher $\dot{V}O_{2peak}$ category values tend to be underestimated.

The few studies comparing end consumer smartwatches found similar yet slightly greater error rates. Previous researchers investigated the $\dot{V}O_{2peak}$ provided by the Garmin Forerunner 920XZ (a preceding model of the smartwatch employed here) and observed a MAPE of 7.3 % in individuals with a mean $\dot{V}O_{2peak}$ of 50.3 ± 8.1 ml·min^{-1}·kg^{-1} using similar testing procedures as in our study [10]. Klepin and colleagues also applied similar testing procedures and found the MAPE for a smartwatch model by Fitbit (Fitbit Charge 2, Fitbit Inc., San Francisco, CA, USA) was 9.1 % with a mean $\dot{V}O_{2peak}$ of 47.6 ml·min^{-1}·kg^{-1}[11]. Our experimental procedures differ from previously performed studies as we include reliability testing of the criterion measure as well. The reliability analysis allows us to compare the error that practitioners should expect when measuring runners twice with the criterion measure (e. g., pre- or post a training period) and when using the smartwatch.

In the given sample, for a runner with a $\dot{V}O_{2peak}$ of 50 ml·min^{-1}·kg^{-1}, the percent variability of the criterion measure is 3.5 %, corresponding to an absolute variability of 1.75 ml·min^{-1}·kg^{-1}. For a runner with a $\dot{V}O_{2peak}$ of 60 ml·min^{-1}·kg^{-1} this variability is 4.0 % (2.4 ml·min^{-1}·kg^{-1}). When employing the criterion measure, any changes of $\dot{V}O_{2peak}$ smaller than 1.75 to 2.4 ml·min^{-1}·kg^{-1} (depending on the level of $\dot{V}O_{2peak}$) should therefore be interpreted cautiously, at least in the given sample and test set-up. In individuals with a $\dot{V}O_{2peak}$ of 45 to 55 ml·min^{-1}·kg^{-1}, variability of the smartwatch and the criterion measure are similar (at least in the given sample) and can be employed interchangeably to assess $\dot{V}O_{2peak}$. In individuals with a $\dot{V}O_{2peak}$ > 55 ml·min^{-1}·kg^{-1} or < 45 ml·min^{-1}·kg^{-1}, the criterion measure shows lower variability than the smartwatch and can therefore better detect smaller changes in $\dot{V}O_{2peak}$.

Usefulness and limitations of $\dot{V}O_{2peak}$ measurement for training in runners

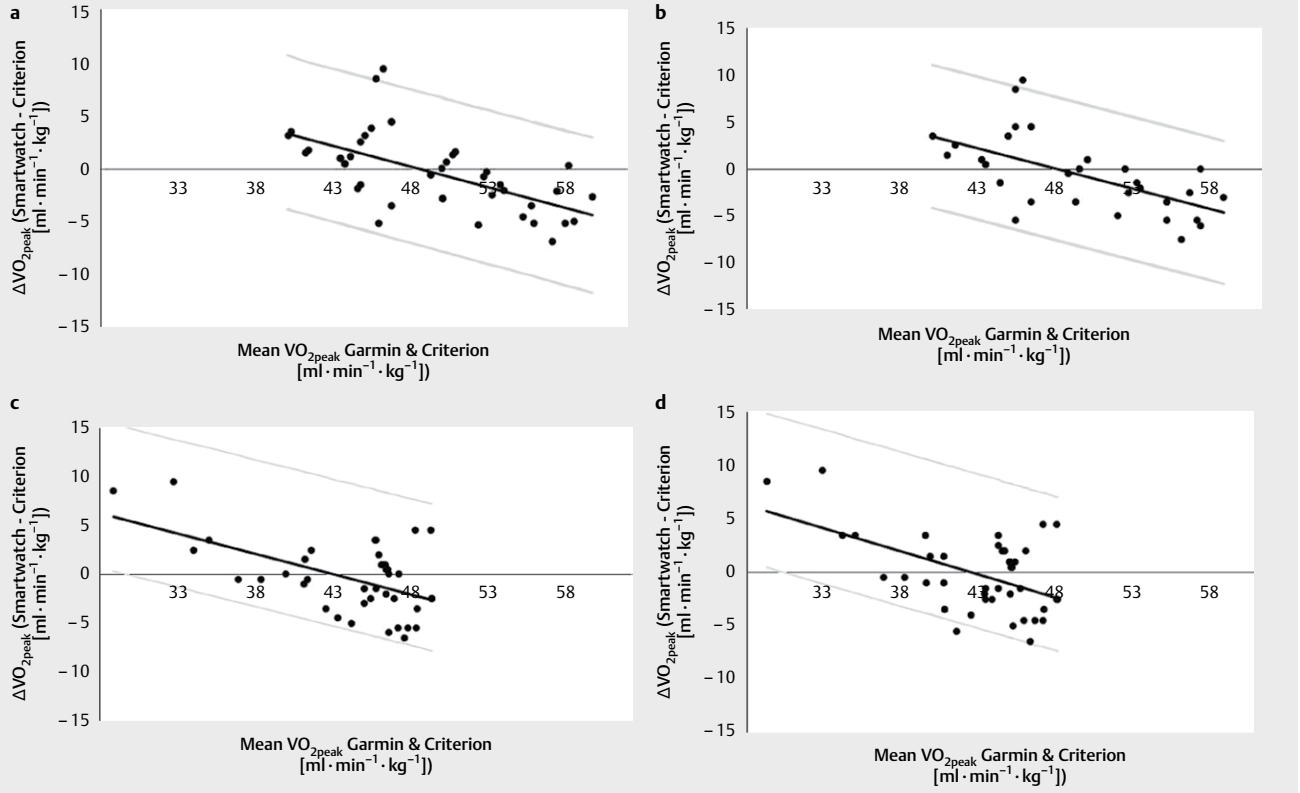
Changes in $\dot{V}O_{2peak}$ allows runners to evaluate the effectiveness of their previous training procedures with regards to maximal oxygen consumption, however in this case the validity of the provided

▶ **Table 1** Descriptive statistics of the main variables obtained during the 1st and 2nd laboratory tests (mean ± SD).

	1st Laboratory test	2nd Laboratory test	Average of laboratory tests
Peak oxygen uptake [ml·min^{-1}·kg^{-1}]	47.5 ± 6.8	49.7 ± 7.3	48.6 ± 7.0
Peak heart rate [bpm]	196 ± 9	193 ± 8	195 ± 8
Peak respiratory exchange ratio	1.14 ± 0.06	1.13 ± 0.26	1.14 ± 0.18
Peak blood lactate concentration [mmol·L^{-1}]	7.0 ± 1.9	6.1 ± 1.5	6.6 ± 1.9
Completed stages on treadmill [n]	7.7 ± 2.4	7.8 ± 2.3	7.5 ± 2.3

▶ **Table 2** Descriptive statistics of variables obtained during the outdoor runs (mean ± SD).

	1st Run	2nd Run	3rd Run	Average of all runs
Duration [s]	2437 ± 608	2456 ± 676	1784 ± 59	2232 ± 614
Distance [km]	6.99 ± 2.17	7.08 ± 2.41	5.89 ± 0.93	6.67 ± 2.01
Mean velocity [m·s^{-1}]	2.86 ± 0.41	2.86 ± 0.60	3.30 ± 0.96	3.00 ± 0.71
Mean heart rate [bpm]	161.3 ± 27.5	161.3 ± 28.8	172.5 ± 36.6	164.9 ± 31.4
Mean heart rate [% of peak heart rate]	82.7 %	82.7 %	88.4 %	84.5 %
Mean ratings of perceived exertion [Borg 6–20 scale]	16 ± 2	16 ± 2	18 ± 2	16 ± 2



▶ **Fig. 2** Bland-Altman plots (mean $\dot{V}O_{2peak}$ of the criterion vs. mean $\dot{V}O_{2peak}$ of the smartwatch) for **a** mean values of 3 outdoor runs, **b** only the first run, **c** only the second run, **d** only the third run.

$\dot{V}O_{2peak}$ values need to be considered for assessing meaningful changes in $\dot{V}O_{2peak}$.

For example, when using the smartwatch derived $\dot{V}O_{2peak}$ measurement, (and based on the present data) a runner with a $\dot{V}O_{2peak}$ of 50 ml·min^{-1}·kg^{-1} will need a change of at least 2 ml·min^{-1}·kg^{-1} to be confident that the displayed change may represent a "true" physiological change and not a measurement error due to low validity. Based on our data, runners with a greater baseline $\dot{V}O_{2peak}$ (> 60 ml·min^{-1}·kg^{-1}) will need a change in $\dot{V}O_{2peak}$ of at least 3.5 ml·min^{-1}·kg^{-1}. When using the present smartwatch model, any smaller change should be interpreted with caution when evaluating the response of $\dot{V}O_{2peak}$ to training.

Based on the miniature design and advanced technology, the smartwatch allows more frequent assessment of $\dot{V}O_{2peak}$ than it would be possible with laboratory measurement such as stationary or portable gas analysis. Among other factors [25, 26] regular (bio-)feedback [27] (e. g., concerning $\dot{V}O_{2peak}$ changes) may ensure a certain level of adherence to training procedures for some runners.

$\dot{V}O_{2peak}$ often also serves as an anchor measurement to prescribe exercise intensity [7]. For example, exercise at an intensity of 40–60 % of $\dot{V}O_{2peak}$ is considered as "moderate," whereas an intensity of 60–80 % of $\dot{V}O_{2peak}$ is considered as "vigorous (hard)" according to the American College of Sports Medicine guidelines for exercise testing and prescription [28]. However large variation in homeostatic perturbations (e. g., oxygen uptake kinetics, blood lactate responses) have been reported across multiple studies for exercise performed within those percentages of $\dot{V}O_{2peak}$ [7]. Consequently, applying fixed percentages of $\dot{V}O_{2peak}$ to define exercise intensity have shortcomings for normalizing between individuals owing to large inter-individual variation in response [7]. Future studies need to further elaborate the individual response to exercise prescribed as fixed percentages of $\dot{V}O_{2peak}$ or whether individual percentages of $\dot{V}O_{2peak}$ are more beneficial to prescribe training procedures.

In summary, while $\dot{V}O_{2peak}$ measurements obtained by a smartwatch might reveal changes in training adaptation (acknowledging that favorable adaptations such as peak cardiac output or mitochondrial oxidative capacity can occur without improvements in $\dot{V}O_{2peak}$ [6]), using $\dot{V}O_{2peak}$ as an anchor measurement to prescribe exercise intensity has limited applicability in guiding training procedures owing to large inter-individual variations in response.

Limitations

We investigated healthy and comparably fit individuals with a $\dot{V}O_{2peak}$ ranging from 38–61 ml·min^{-1}·kg^{-1} and did not include participants with higher or lower cardiorespiratory fitness. Cautious interpretation is warranted when transferring our results to other populations, e. g., cardiac patients with altered heart dimension and/or function or individuals with exceptional cardiac dimensions such as elite athletes. Also, our set-up was designed for runners and not for cycling or other sports; therefore we advise to test the validity of $\dot{V}O_{2peak}$ measurements in different sports involving different movement patterns than running. Additionally, future studies might evaluate whether more running sessions alter the validity of

the provided $\dot{V}O_{2peak}$ measurements. Furthermore, future studies should also evaluate if the validity is affected by performing runs of different duration or intensity and in different weather and environmental conditions (e. g., frequent strong headwind or running on sand). The reason for less valid $\dot{V}O_{2peak}$ estimations of the smartwatch > 55 ml·min^{-1}·kg^{-1} or < 45 ml·min^{-1}·kg^{-1} are currently elusive and need further investigation.

As our aim was to test the validity of a smartwatch to estimate $\dot{V}O_{2peak}$ for end consumer purposes, we did not enter each runner's peak heart rate into the smartwatch software since estimations with formulas are subject to error [29] and recreational runners often do not know their actual peak heart rate. Therefore, the present results might be different when entering a runner's true peak heart rate into the software. Additionally, the results may also differ when runners wear a heart rate belt that may assess the heart rate more accurately than the optical heart rate monitor, especially at higher running velocity.

Conclusions

In the given group of runners as well as the applied testing procedures and within the $\dot{V}O_{2peak}$ range of 45 and 55 ml·min^{-1}·kg^{-1}, the mean absolute percentage error when validating against the criterion measure is 4.1 %. The criterion measure revealed a coefficient of variation of 3.5 % in this range of $\dot{V}O_{2peak}$.

$\dot{V}O_{2peak}$ measurement with the smartwatch in runners with lower (< 45 ml·min^{-1}·kg^{-1}) or higher (> 55 ml·min^{-1}·kg^{-1}) $\dot{V}O_{2peak}$ should be judged cautiously due to higher error rates between the smartwatch and the criterion measure.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Zinner C, Sperlich B, Wahl P et al. Classification of selected cardiopulmonary variables of elite athletes of different age, gender, and disciplines during incremental exercise testing. SpringerPlus 2015; 4: 544

[2] Duking P, Holmberg HC, Kunz P et al. Intra-individual physiological response of recreational runners to different training mesocycles: A randomized cross-over study. Eur J Appl Physiol 2020; 120: 2705–2713

[3] Jones AM, Kirby BS, Clark IE et al. Physiological demands of running at 2-hour marathon race pace. J Appl Physiol (1985) 2020; 120: 2705–2713

[4] Milanovic Z, Sporis G, Weston M. Effectiveness of high-intensity interval training (HIT) and continuous endurance training for VO_2max improvements: A systematic review and meta-analysis of controlled trials. Sports Med 2015; 45: 1469–1481

[5] Daniels JT, Yarbrough RA, Foster C. Changes in VO_2 max and running performance with training. Eur J Appl Physiol Occup Physiol 1978; 39: 249–254

[6] Martin-Rincon M, Calbet JAL. Progress update and challenges on V. O_2max testing and interpretation. Front Physiol 2020; 11: 1070

[7] Jamnick NA, Pettitt RW, Granata C et al. An examination and critique of current methods to determine exercise intensity. Sports Med 2020; 50: 1729–1756

[8] Zinner C, Olstad DS, Sperlich B. Mesocycles with different training intensity distribution in recreational runners. Med Sci Sports Exerc 2018; 50: 1641–1648

[9] De Brabandere A, Op De Beeck T, Schutte KH et al. Data fusion of body-worn accelerometers and heart rate to predict VO_2max during submaximal running. PLoS One 2018; 13: e0199509

[10] Passler S, Bohrer J, Blochinger L et al. Validity of wrist-worn activity trackers for estimating VO2max and energy expenditure. Int J Environ Res Public Health 2019; 16: 3037

[11] Klepin K, Wing D, Higgins M et al. Validity of cardiorespiratory fitness measured with fitbit compared to VO2max. Med Sci Sports Exerc 2019; 51: 2251–2256

[12] Düking P, Fuss FK, Holmberg HC et al. Recommendations for assessment of the reliability, sensitivity, and validity of data provided by wearable sensors designed for monitoring physical activity. JMIR MHealth UHealth 2018; 6: e102

[13] Düking P, Hotho A, Holmberg HC et al. Comparison of non-invasive individual monitoring of the training and health of athletes with commercially available wearable technologies. Front Physiol 2016; 7: 71

[14] Van Hooren B, Goudsmit J, Restrepo J et al. Real-time feedback by wearables in running: current approaches, challenges and suggestions for improvements. J Sports Sci 2020; 38: 214–230

[15] Harriss DJ, MacSween A, Atkinson G. Ethical standards in sport and exercise science research: 2020 update. Int J Sports Med 2019; 40: 813–817

[16] Atkinson G, Williamson P, Batterham AM. Issues in the determination of 'responders' and 'non-responders' in physiological research. Exp Physiol 2019; 104: 1215–1225

[17] Garmin Ltd Forerunner (r) 245/245 Music Benutzerhandbuch 2019

[18] Schaun GZ. The maximal oxygen uptake verification phase: A light at the end of the tunnel? Sports Med Open 2017; 3: 44

[19] Borg G. Perceived exertion as an indicator of somatic stress. Scand J Rehabil Med 1970; 2: 92–98

[20] Macfarlane DJ, Wong P. Validity, reliability and stability of the portable Cortex Metamax 3B gas analysis system. Eur J Appl Physiol 2012; 112: 2539–2547

[21] Firstbeat Technologies Ltd Automated Fitness Level (VO2max) Estimation with Heart Rate and Speed Data 2014, available at internet at https://www.semanticscholar.org/paper/Automated-Fitness-Level-(-VO-2-max-)-Estimation-and/f5a83536ffd3948ca-522f86e514835912853e29d

[22] Winkert K, Kirsten J, Dreyhaupt J et al. The COSMED K5 in breath-by-breath and mixing chamber mode at low to high intensities. Med Sci Sports Exerc 2020; 52: 1153–1162

[23] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016; 15: 155–163

[24] Lee JM, Kim Y, Welk GJ. Validity of consumer-based physical activity monitors. Med Sci Sports Exerc 2014; 46: 1840–1848

[25] McArthur D, Dumas A, Woodend K et al. Factors influencing adherence to regular exercise in middle-aged women: a qualitative study to inform clinical practice. BMC Womens Health 2014; 14: 49. 26

[26] Robison JI, Rogers MA. Adherence to exercise programmes. Recommendations. Sports Med 1994; 17: 39–52

[27] Lyons EJ, Lewis ZH, Mayrsohn BG et al. Behavior change techniques implemented in electronic lifestyle activity monitors: a systematic content analysis. J Med Internet Res 2014; 16: e192

[28] American Colleague of Sports Medicine ACSM's Guidelines For Exercise Testing And Prescription. Philadelphia: Lippincott Williams & Wilkins; 2013

[29] Sarzynski MA, Rankinen T, Earnest CP et al. Measured maximal heart rates compared to commonly used age-based prediction equations in the Heritage Family Study. Am J Hum Biol 2013; 25: 695–701