Pilot study assessing the influence of skin type on the heart rate measurements obtained by photoplethysmography with the Apple Watch

[1]BORJA SAÑUDO, [1]MOISÉS DE HOYO, [2]ALEJANDRO MUÑOZ-LÓPEZ, [3]JOHN PERRY,
[4]GRANT ABT

Running Head: Skin Type and photoplethysmography

[1]*Department of Physical Education and Sport, University of Seville, Seville, Spain*

[2]*Science to Improve, Seville, Spain*

[3]*Department of Psychology, Mary Immaculate College, Ireland.*

[4]*Department of Sport, Health and Exercise Science, University of Hull, United Kingdom*

Conflicts of interest: none

Source of support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Corresponding Author:
Borja Sañudo, PhD
https://orcid.org/0000-0002-9969-9573
Physical Education and Sports Department
University of Seville
Pirotécnia s/n, N-41012
Seville, Spain
Tel: ++34 652 387090
Email: bsancor@us.es

ABSTRACT

Photoplethysmographic imaging (PPG) is currently used to measure heart rate (HR) and the accuracy of PPG can be influenced by pigmentation of the skin; however, the effects of skin color-related artifacts on PPG during exercise remain unclear. This study aimed to assess the agreement between the Apple Watch photoplethysmography sensor and a criterion, for measuring heart rate across a range of intensities during exercise and to determine the influence of skin type on the accuracy of the measure. Forty-five males (20-43 y) completed the Fitzpatrick Skin Scale and were classified into three different skin type groups: a) types II (n=15), III (n=15) and IV (n=15). Participants performed a graded incremental cycle-ergometer test while simultaneously wearing the Apple Watch and a Polar monitor as a criterion measure. Data from both devices were collected in 5-s epochs. Correlations between devices were very good (0.96-0.99 [95%CI: 0.94 to 0.99]). Significant differences were observed between skin types II and III when the intensity of the exercise was increased, albeit with trivial to small effect sizes (ES: 0.05 to 0.28). All significant differences corresponded to <2% of relative difference between both devices. Bland-Altman analyses showed a trivial but systematic underestimation of HR in the Apple Watch compared to Polar for all skin types during exercise. In conclusion, the Apple Watch accurately measures HR when cycling at different intensities and certain types of skin seem not to influence these measures, which may have important implications for controlling the intensity of exercise.

Key Terms: Heart rate; agreement; wearable sensors; exercise; skin type

1. INTRODUCTION

Heart rate (HR) is commonly used to monitor exercise intensity and therefore accurate measures are important to provide individuals with precise estimates of cardiovascular-based exercise intensity for safe and effective workouts. For many years, noninvasive techniques for monitoring HR, such as portable electrocardiography (ECG) monitors, have been analyzed. Most of these devices detect HR via a chest strap (e.g., Polar®), which has been shown to be both a valid and a reliable method for determining HR during rest and exercise [1,2]. However, some individuals are unable (e.g., sensitive skin) or unwilling (e.g., the attachment of the electrode may be troublesome and the strap needs to be worn on the skin and kept wet for accurate signal detection) to use these methods [3]. Therefore, due to the problems experienced when fitting the strap and the discomfort reported by many during exercise, especially when worn for extended periods [4], the use of other alternatives has been recommended [5].

Optical methods, such as photoplethysmographic imaging (PPG), are also widely used and have been investigated in recent years as an alternative to overcome the limitations of traditional methods [5]. Numerous wearable activity trackers, such as the Apple Watch®, incorporate optical LED sensors to non-invasively detect changes in the light intensity with respect to the change in volume of blood flow and thus measure HR [5,6]. This technology consists of a light source to illuminate the skin tissue, and a photo-detector to measure small variations in light intensity associated with changes in perfusion in peripheral blood vessels [3]. The simplicity and easy accessibility of PPG has meant that many use continuous HR monitoring to control exercise intensity without being aware of the implications this may have on their performance or health.

PPG may be affected by several external factors such as anatomical placement (pressure between a probe and the skin), environmental noise (ambient light), sweat and especially motion artifacts, as signals are very sensitive to small changes in sensor position which are considered an important obstacle when computing HR from PPG [4,6,7]. Also, the accuracy of PPG can be dependent on the type and intensity of the exercise [8]. Further, the sensitivity of the sensor may be influenced by pigmentation of the skin [9]. However, while many signal processing techniques have been proposed to remove motion artifacts during exercise, the effects of skin color-related artifacts on PPG during this practice remain unclear [4].

Preliminary studies seem to indicate that variability caused by the amount of melanin may affect characteristics of PPG signals [4,9,10] and while green wavelength, which is the one used in the Apple Watch, bring greater signal resolution during exercise [11], evidence of the accuracy of these sensors when measuring HR in people of varied skin pigment is scarce [4]. Thus, in order to correctly monitor exercise intensity (i.e., HR) there is a clear consensus on the importance of validation studies incorporating separate analyses specific to subject skin color [8]. Therefore, the aims of this pilot study were to (a) determine the validity of the Apple Watch PPG sensor when measuring HR across a range of exercise intensities in reference to the Polar device criterion measure which employed a chest strap based technology and has been shown to be highly correlated to ECG [1,2] and (b) show some preliminary data on the influence of skin type on the accuracy of the measure. We hypothesized that at certain intensities (i.e., high intensities) the light reflectance would be different across skin types.

2. MATERIALS AND METHODS

2.1 Participants

Forty-five healthy males were recruited to this study. Participants were between the ages of 20 and 43 years (24 ± 4 y; body mass: 72.2 ± 5.8 kg; stature: 1.77 ± 0.05 m) and engaged in physical activities at least three times per week. The study was conducted at the Sport Science Lab at the University of Seville. Participants were recruited by visiting scheduled classes and asking for volunteers to complete the test. Participants were eligible if they were between 18 and 45 years of age, and did not have any history of injury or disease (e.g., peripheral circulatory failure) that would prevent them from safely performing the study protocol. All participants refrained from smoking, caffeine intake, alcohol consumption and extreme exercise for 12 h before the experiment to minimize effects that could affect blood flow. The study was approved by the Institutional Review Board of the University of Seville and after being informed of the purpose, procedures, benefits and risks of the study, written informed consent was obtained from each participant.

2.2. Procedures

As the skin perfusion also changes with environment, the experiments were performed in the laboratory under temperature-controlled conditions. After arrival at the laboratory, participants were required to complete the Fitzpatrick Skin Scale [12] where the range consisted from type I = high photosensitivity to type VI = low photosensitivity. After comparing with a photograph of each subject’s forearm, participants were then classified by an assistant not involved in the study into three different Fitzpatrick skin type groups, ranging from type II to type IV (as no participants with type I, type V or type

VI skin photosensitivity participated in the study). All groups were equally sized with 15 participants in each.

The Apple Watch was placed on the forearm approximately 2 cm from the wrist bone according to the manufacturer's specifications. The criterion measure of HR was measured via a HR receiver (Polar RS800CX monitor, Polar Electro OY, Kempele, Finland) that was placed on the left wrist and an accompanying chest strap that was applied as per manufacturer's instructions. This device has been shown to be a valid gold-standard measure of mobile HR monitoring technology when compared with ECG measurements during exercise [1,2]. Body mass and stature were assessed and then participants were kept in a quiet room in a seated position for 10 min while their resting HR was measured. Then, each participant immediately started an incremental graded exercise test on a cycle ergometer.

The incremental graded exercise test started after a standardized warm-up consisting of 5 min of pedaling on a cycle ergometer (Ergoselect 200, Ergoline GmbH, Bitz, Germany) at a load of 50 W. Then participants performed a maximal graded exercise test at an initial load of 50 W (cadence of 60 rpm) that increased by 25 W every one min until exhaustion. Data from the Polar and Apple Watch devices were collected in 5-s epochs by reading each HR value from the watch face. HR values were independently registered by two assistants. These values were used to calculate the mean HR over each minute while performing the incremental protocol. For each group, HR was divided in percentage zones from each individual peak HR to compare relative zones between both devices (Zone 1 = 0-59%, Zone 2 = 60-69%, Zone 3 = 70-79%, Zone 4 80-89% and Zone 5 = 90-100%).

2.3. Data analysis

Data are represented as mean (SD) for each device and phototype. Prior to assessing the relative HR within each zone from both devices for each group, normality was assessed using the Kolmogorov-Smirnov test. All data violated the assumption of normality and therefore a non-parametric Mann-Whitney U test was used to assess possible mean differences between HR zones. A P value of < 0.05 was used to determine whether the possible differences were statistically significant or not. In addition, the standard error of the mean was calculated (SSE). To assess the magnitude of the differences, Cohen's *d* effect size (ES) was calculated by dividing the pooled standard deviation by the mean differences between both devices in each HR zone. The following magnitudes were used to interpret the ES: trivial effect: <0.20, small effect: from 0.20 to 0.59, moderate effect: from 0.60 to 1.19, large effect: from 1.20 to 1.99, very large effect: > 1.99 [13].

To determine the agreement between both instruments, two separate analyses were conducted. First, Pearson product moment correlations and 95% confidence intervals (CI) were used for each pair of HR data for each group. Prior to any plots analysis, data were log transformed to reduce non-uniformity associated errors. The following magnitudes were used to interpret the correlations: very poor (r = 0.45 to 0.69), poor (r = 0.70 to 0.84), good (r = 0.85 to 0.94), very good (r = 0.95 to 0.994) and excellent (r ≥ 0.995) [14]. Second, to calculate absolute systematic bias, Bland-Altman plots for repeated measures were used for each group, together with the corresponding 95%

limits of agreement (LoA) following the guidance of Bland & Altman [15], using calculations provided by Zou [16]. For all measures, the true value was assumed to vary. Finally, the coefficient of correlation (r^2) of the plots were calculated to assess either if bias was constantly along all the data ($r^2 < 0.1$) or tended to overestimate lower or higher heart rates [17]. All calculations were conducted using SPSS (version 22.0, Chicaco, IL).

3. RESULTS

HR values together with the SEE and ES for each skin type at rest (Table 1) and during the graded incremental exercise test (Table 2) are reported for both devices. At rest, there were no significant differences between both devices for any skin type (trivial ES).

During exercise, type II participants showed significant differences between Apple Watch and Polar in zones: 70-79% (135 ± 10 vs 138 ± 10 respectively, small ES), 80-89% (154 ± 12 vs 157 ± 11 respectively, small ES) and 90-100% (175 ± 12 vs 178 ± 12 respectively, trivial ES). Type III showed significant differences between devices within the same zones: 70-79% (138 ± 8 vs 140 ± 8 respectively, small ES), 80-89% (158 ± 9 vs 160 ± 8 respectively, small ES) and 90-100% (176 ± 9 vs 178 ± 9 respectively, small ES), while non-significant differences were observed for Type IV in any zone. All significant differences found corresponded to <2% of relative difference between both devices (from 1.2 to 2.1% [CI: 1.1 to 1.8]).

Figures 1 and 2 displays the Pearson product-moment correlation coefficient (r) for the Apple Watch showing excellent correlations with the criterion measure during exercise (all r=0.99 [0.99-0.99 CI], p<0.001). Good to excellent correlations were also observed during the rest condition in all groups (type II= 0.98 [0.97-0.99 CI], Type III= 0.96 [0.94-0.98 CI], type IV= 0.98 [0.98-0.99 CI]) (Figure 2).

Bland-Altman analyses (mean difference and limits of agreement) are presented in Figure 3.

There was a proportional systematic bias in the recorded HR between both devices for all the skin types during the exercise condition (mean bias [95%LoA]): type II= -2(-8 to 5) beats·min^{-1}, type III= -2(-8 to 4) beats·min^{-1}, type IV= -1(-6 to 4) beats·min^{-1}. In the resting condition, participants in type II exhibited a mean bias of 0 (-5 to 4), type III= 0 (-5 to 5) and type IV= 0 (-5 to 4).

4. DISCUSSION

While previous studies investigated the accuracy of wrist wearable technologies for estimating HR at different intensities [5,18,19], to our knowledge this is the first study to examine how well the Apple Watch wrist-worn device agrees with a criterion measure of HR during rest and cycling at different intensities while examining the influence of a range of skin types on this agreement. The results obtained in the current study suggest that the Apple Watch agrees well with the criterion measure and therefore fulfills published criteria for HR measurement provided in previous research [20]: a) A correlation r=0.90 or greater between the test device and the criterion measure; b) A mean bias less than 3 beats·min^{-1}; c) A standard error less than 5 beats·min^{-1}.

Our results showed good to excellent correlations during exercise (all r>0.9) and mean bias <2%. Therefore, and following previous recommendations coming from the validation of consumer devices for accurate HR measurement [8] we can state that the Apple Watch, which continually measures HR using PPG, can be used during a maximal graded exercise test on a cycle ergometer. However, the question that arises is whether the Apple Watch is still accurate depending on the skin type of the participants. In this sense, we observed good to excellent correlations in all groups (r > 0.93) between the devices with a mean bias <2 beats·min^{-1} and an absolute difference from criterion measurements of <2%.

When analyzing the data as a whole, regardless of the skin type, our results are in accordance with Wallen et al. [19] who also examined the accuracy of different wrist-worn devices, including the Apple Watch, to measure HR. These authors reported that the devices were within 1–9% of reference estimates. However, they also reported that all devices underestimated HR. In the same line, Dooley et al. [18] recently reported that the magnitude of errors across all intensities (treadmill exercise) for the Apple Watch were between 1.1%-6.7%. In the current study the mean absolute percentage error observed was <2% at all intensities. These discrepancies can be attributable to numerous factors such as the mode of exercise, intensity, and participant characteristics. If we focus on intensity, Jo et al. [8] reported that the performance of another wrist-worn device (Fitbit Charge HR) was poor during low intensity cycling (60 W) and usually, the accuracies were reduced with increasing exercise intensity. In fact, it was recently reported that the correlation for the Apple Watch decreased at high intensities [18,19], which contrasts slightly with the results of our study.

The mode of exercise may also reduce the correlation between both measures [4]. Thus, Wallen et al. [20] revealed that HR measurement error tends to differ between treadmill and cycle protocols and recently, Shcherbina et al. [21] also reported the lowest error in measuring HR for the cycle ergometer task of 1.8% (0.9%–2.7%). These figures are in agreement with the ones reported in our study - 1.59% (0.93%–2.09%). This could be an explanation for our results since this tendency showing greater error when speeds were increased was also observed by Lee & Gorelick [22] in the validation study of a different smart watch and they suggested that this could be due to the greater disturbances by the movement, the sensitivity of the device or even by the skin type of the participant being studied. However, due to the large number of variables affecting PPG, it is not practical to include all variables in a single study. For instance, the movement of the Apple Watch was not varied in our study. This had the benefit of not introducing a potentially confounding variable, but it needs to be considered that motion artifacts can have powerful effects on the efficacy of the HR measurement [9].

Therefore, regarding the skin type, the Apple Watch was not significantly different from the Polar HR monitor during baseline (all $P>0.873$). On the basis of the Bland-Altman analysis there was a systematic bias in the recorded HR between both devices for all the phototypes during the exercise condition (bias ≤2 beats·min^{-1} and the 95% limits of agreement: -8 to 5 beats·min^{-1}) which is consistent with Wallen et al. [20] who previously validated the Apple Watch showing a very good correlation ($r = 0.95$) with ECG and a small mean bias of -1 beats·min^{-1}. Also during exercise, Spierer et al. [4] indicated that some skin types could produce more error in some wrist-worn devices (Mio Alpha). In the same line, Fallow et al. [11] demonstrated that a dark skin type (type V) attenuated the signal in comparison with other skin types. It is known that melanin can absorb light and thus attenuate the incident light wavelength [11]. However, while a dark pigmentation was suggested to lead to a worse light reflection [11], in the current study non-significant differences between both devices were observed for the darker skin type analysed (type IV). This contrasts with Wallen et al. [20] who reported statistical differences between correlations for HR based on skin colour (skin Type >IV was statistically different to skin Type <IV). In any case, we observed diminished performance when cycling at higher intensities (i.e. >70-79%) in participants with skin type II and III. The measurement was accurate monitoring HR even with increasing physical exertion, although some differences existed between types II and III, especially in the higher intensity zones. An alternative explanation to these discrepancies can be attributed to changes in the position of the sensor (the proximity of the device on the wrist) derived from an isometric muscle contraction while holding on to handlebars while cycling. It was suggested that in this activity, the fluctuations associated with muscle actions might affect PPG signals [23].

There are a number of limitations that have to be considered when interpreting the results. The first refers to the sample, where only healthy, relatively young (20–43 years) individuals within the normal range of body composition were included, which could limit the generalization of the results to other population groups (e.g. older adults). Second, and despite having the same number of participants per group, we did not find sufficient number of people classified as skin type I, V and VI, although this seems logical taking into account the characteristics of the population in which the study was developed.

Despite these limitations, the current study shows a strong agreement between the Apple Watch and the criterion measure when exercising on a cycle ergometer at different intensities, which together with a low systematic bias ensure that both devices may be used inter-changeably for accurate HR measurements. Moreover, some preliminary results on the effect of skin type suggest that the skin types analyzed have no influence on the heart rate values obtained.

PERSPECTIVES

Heart rate monitors are widely used to control exercise intensity; however, many athletes complain about having to use the chest bands. The results of this study provide scientists, coach's and clinicians the error measurement of the Apple Watch when cycling at different intensities. The study used a novel approach to measure accuracy of this device for HR at specific bouts of exercise intensities according to the skin pigmentation, which may be relevant in the sport medicine area when controlling or prescribing physical activity by means of this PPG sensor. While the practicality of the tested sensor has to be examined in a future studies, especially in a real-life setting or with respect to different activities, we show important findings since this wrist-worn device utilizing PPG offers both medical staff and performance coaches a valid method to monitor HR while exercising on a cycle ergometer, which is essential to control the exercise intensity.

Funding

This study had no funding support.

Compliance with Ethical Standards

Authors declare that there are no conflicts of interest.

Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed Consent Informed consent was obtained from all individual participants included in the study.

REFERENCES

1. Giles D, Draper N and Neil W (2016) Validity of the Polar V800 heart rate monitor to measure RR intervals at rest. *Eur J Appl Physiol* 116(3):563-71. Http://doi.org/ 10.1007/s00421-015-3303-9.
2. Barbosa MP, da Silva NT, de Azevedo FM, Pastre CM and Vanderlei LC (2016) Comparison of Polar® RS800G3™ heart rate monitor with Polar® S810i™ and electrocardiogram to obtain the series of RR intervals and analysis of heart rate variability at rest. *Clin Physiol Funct Imaging* 36(2):112-7. Http://doi.org/10.1111/cpf.12203.
3. Maeda Y, Sekine M and Tamura T (2011). Relationship between measurement site and motion artifacts in wearable reflected photoplethysmography. *J Med Syst* 35(5):969-76. Http://doi.org/10.1007/s10916-010-9505-0.
4. Spierer DK, Rosen Z, Litman LL and Fujii K (2015) Validation of photoplethysmography as a method to detect heart rate during rest and exercise. *J Med Eng Technol* 39(5):264-71. Http://doi.org/10.3109/03091902.2015.1047536.
5. Stahl SE, An HS, Dinkel DM, Noble JM and Lee JM (2016) How accurate are the wrist-based

heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport Exerc Med* 2(1):e000106.

6. Madhan Mohan P, Nagarajan V and Vignesh JC (2017) Spot measurement of heart rate based on morphology of PhotoPlethysmoGraphic (PPG) signals. *J Med Eng Technol* 41(2):87-96.
7. Zhu S, Tan K, Zhang X, Liu Z and Liu B (2015) MICROST: A mixed approach for heart rate monitoring during intensive physical exercise using wrist-type PPG Signals. *Conf Proc IEEE Eng Med Biol Soc* 2347-50. Http://doi.org/10.1109/EMBC.2015.7318864.
8. Jo E, Lewis K, Directo D, Kim MJ and Dolezal BA (2016) Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking. J Sports Sci Med 15(3):540-547.
9. Allen J (2007) Photoplethysmography and its application in clinical physiological measurement. *Physiol Meas* 28(3):R1-39.
10. Butler MJ, Crowe JA, Hayes-Gill BR and Rodmell PI (2016) Motion limitations of non-contact photoplethysmography due to the optical and topological properties of skin. *Physiol Meas* 37(5):N27-37. Http://doi.org/10.1088/0967-3334/37/5/N27
11. Fallow BA, Tarumi T and Tanaka H (2013) Influence of skin type and wavelength on light wave reflectance. J Clin Monit Comput 27(3):313-7. Http://doi.org/10.1007/s10877-013-9436-7.
12. Fitzpatrick TB (1988) The validity and practicality of sun-reactive skin types I through VI. *Arch Dermatol* 124(6):869-71.
13. Hopkins WG, Marshall SW, Batterham AM and Hanin J (2009) Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc* 41(1):3-13. Http://doi.org/10.1249/MSS.0b013e31818cb278
14. Hopkins WG (2016) Validity thresholds and error rates for test measures used to assess individuals. In Proc. 21st Annu. Congress of the European College of Sport Science, Vienna, Austria.

15. Bland JM and Altman DG (2007) Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat* 17(4):571-82.
16. Zou GY (2013) Confidence interval estimation for the Bland-Altman limits of agreement with multiple observations per individual. *Stat Methods Med Res* 22(6):630-42. Http://doi.org/10.1177/0962280211402548
17. Atkinson G and Nevill AM (1998) Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 26(4):217-38.
18. Dooley EE, Golaszewski NM and Bartholomew JB (2017) Estimating Accuracy at Exercise Intensities: A Comparative Study of Self-Monitoring Heart Rate and Physical Activity Wearable Devices. *JMIR Mhealth Uhealth* 5(3):e34. Http://doi.org/10.2196/mhealth.7043.
19. Khushhal A, Nichols S, Evans W, et al. Validity and Reliability of the Apple Watch for Measuring Heart Rate During Exercise. *Sports Med Int Open* 2017*;*1(6):E206-E11. http://doi.org/10.1055/s-0043-120195.
20. Wallen MP, Gomersall SR, Keating SE, Wisløff U and Coombes JS (2016) Accuracy of Heart Rate Watches: Implications for Weight Management. *PLoS One* 11(5):e0154420. Http://doi.org/10.1371/journal.pone.0154420.
21. Shcherbina A, Mattsson CM, Waggott D, et al (2017) Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *J Pers Med* 7(2). pii: E3. Http://doi.org/10.3390/jpm7020003.
22. Lee CM and Gorelick M (2011) Validity of the Smarthealth watch to measuring heart rate during rest and activity. *Meas Phys Educ Exerc Sci* 15(1):18-25.
23. Kamshilin AA, Mamontov OV, Koval VT, Zayats GA and Romashko RV (2015) Influence of a skin status on the light interaction with dermis. *Biomed Opt Express* 6(11):4326-34. Http://doi.org/10.1364/BOE.6.004326.

Figure legends

***Figure 1**. Exercise data correlations plots. HR= heart rate. A= Type II; B= Type III; C= Type IV*

***Figure 2.** Resting data correlations plots. HR= heart rate. A= Type II; B= Type III; C= Type IV*

Figure 3. Bland-Altman plot showing the mean bias and 95% limits of agreement (with 95% confidence intervals) for the absolute differences in heart rate (% HR_{max}) in participants with skin type II (Figure 3a), III (Figure 3c) and IV (Figure 3e) and the relative differences in heart rate (% HR_{max}) in participants with skin type II (Figure 3b), III (Figure 3d) and IV (Figure 3f).