RESEARCH ARTICLE **Open Access**

A Comparison of the Accuracy and Reliability of the Wahoo KICKR and SRM Power Meter

Matthew W. Hoon[1], Scott W. Michael[2]; Raymond L. Patton[2]; Phillip G. Chapman[1]; Jose L. Areta[3]

Abstract

The Wahoo KICKR cycling trainer is a new direct-drive electromagnetically braked bike-trainer that allows cyclists to use their own bicycles as ergometer. It is purported to provide ±3% accuracy in power, despite costing considerably less than other cycling ergometers. The purpose of this study was to assess the accuracy and reliability of several KICKR units against the more established SRM power meter using a first-principles based dynamic calibration rig (CALRIG). Five KICKRs and one SRM unit were assessed by a CALRIG-driven incremental test. Following a 15 min warm-up and 'calibration' as per manufacturer instructions, power was increased (starting at 50 W) by 50 W every 2 min up to 400 W. Each unit was tested twice non-consecutively, in random order. Data was recorded at 1 Hz, with the last 10 s of each stage being averaged for analysis. The mean error (%) and coefficient of determination (R^2) versus CALRIG; as well as the change in mean error and Typical Error of Measurement (TEM) (expressed as a % coefficient of variation) between trials was calculated for each device. The mean error across all KICKR units was -1.5% (range: -3.1% to 0.0%) compared to -1.6% reported by the SRM. R^2 >0.999 was found for all KICKR units and SRM compared to the CALRIG. The mean TEM for the KICKRs was 1.5% (range: 1.1% to 1.9%), whereas the SRM reported 0.7%. For test-retest reproducibility, two KICKRs had statistically significant changes in mean error, with an average 1.3% change across all KICKRs. Comparatively, the SRM reported a 0.4% change between trials. The Wahoo KICKR trainer measures power to a similar level of accuracy to the more reputable SRM power meter during an incremental exercise test. Although not as reproducible, the KICKR still demonstrates an acceptable level of reliability for assessing cycling performance.

Keywords: Cycling, Ramp, Testing, Validity

✉ **Contact email:** matthew.hoon@acu.edu.au (MW. Hoon)

[1]*School of Exercise Science, Australian Catholic University, Strathfield, Australia.*
[2]*Discipline of Exercise and Sports Science, University of Sydney, Lidcombe, Australia*
[3]*Department of Physical Performance, Norwegian School of Sport Sciences, Oslo, Norway*

Received: 10 May 2016. Accepted. 07 Sept 2016.

Introduction

The ability to measure power output accurately allows sport scientists and researchers to assess human cycling performance. This has important implications for a range of purposes, including the determination of athlete training data, performance analysis, and assessing the effect of a particular intervention. Given that meaningful performance gains in competitive sport are typically very small (Hopkins, Hawley, & Burke, 1999), a high level of precision and accuracy is required of power meters, particularly for use in scientific studies.

A new power measuring device, the KICKR (Wahoo Fitness, Atlanta, USA) is a bicycle trainer which may be suitable for use in scientific investigations. The KICKR unit is a computer controlled, electronically-braked system built around a 12.5 lb (5.7 kg) flywheel, which is connected to the drivetrain of a bicycle (Figure 1), and has a manufacturer reported accuracy of ±3% up to 1550 W (http://eu.wahoofitness.com/devices/kickr.html; accessed on 08/02/16). One of the major advantages of the KICKR is that it may be attached to a majority of bicycles to become a controllable cycling ergometer, at a significantly lower cost than traditional dedicated ergometers. Further, participants are able utilize their own bicycles to improve familiarisation and reduce variations in setup. In addition, the KICKR is portable and more easily affixed to different bicycles than other established power meters such as the crank-based SRM (Jülich, Welldorf, Germany).

To date, several power measuring devices have been assessed for use in scientific investigations, such as the Fortius (Bertucci, 2012), Axiom (Bertucci, Duc, Villerius, & Grappe, 2005), Velotron (Abbiss, Quod, Levin, Martin, & Laursen, 2009), Powertap Hub (Bertucci, Duc, Villerius, Pernin, & Grappe, 2005), Wattbike (Hopker, Myers, Jobson, Bruce, & Passfield, 2010), LeMond Revolution (Novak, Stevens, & Dascombe, 2015); and most notably, the SRM (Gardner et al., 2004). Few of these, however, have directly validated against a dynamic calibration rig, the most accurate method of assessment (Hopkins, Schabort, & Hawley, 2001; Paton & Hopkins, 2001).

Currently, only one study has tested the accuracy of the KICKR unit, with the authors concluding that the KICKR demonstrated an acceptable level of accuracy for "training, performance assessment and talent identification" purposes (Zadow, Kitic, Wu, Smith, & Fell, 2016). This investigation, however, was limited to only one KICKR unit, so no assessment of between unit variability was available. Additionally, the reliability of the KICKR was not examined. Therefore, the aim of the current investigation was to: i) assess the accuracy and reliability of several KICKR units and ii) compare it to the more established SRM power meter, when tested against a first principles dynamic calibration device.

Figure 1. a) The attachment of the KICKR to the drivetrain of a bicycle; & b) Superior view. 1: Flywheel; 2: Belt drive; 3: Cassette.

Materials and methods

Material

Five KICKR units were compared against a crank-based power meter (SRM, Jülich, Welldorf, Germany) and a custom-built dynamic calibration rig (CALRIG) with a maximum power-output of 400 W. The testing system comprised of a dedicated alloy road bike fitted with an SRM, mounted sequentially on each KICKR unit in subsequent tests. The SRM was calibrated by the manufacturer a week before testing began. The precision of the SRM was further verified in our laboratory following validated procedures (Wooles, Robinson, & Keen, 2005). The CALRIG's motor provided rotational drive to the crank-set via a universal linkage attached to the non-drive side pedal spindle of the SRM. The rig was fitted with a force transducer (XTran Load Cell S1W, Applied Measurement, Sydney, Australia) to calculate reaction torque, and an optical sensor to determine angular velocity. The transducer was calibrated using a range of known weights (3 point calibration) before the commencement of the study. Data was recorded through PC serial input and power calculated as: reaction torque x angular velocity, at a rate of 1 Hz.

All data from the SRM were recorded by a PC7 head-unit (SRM, Jülich, Welldorf, Germany) at a frequency of 1 Hz. The control of the KICKR and data recording was managed by a personal computer running TrainerRoad software (v2.7.2, TrainerRoad, Nevada, USA), which was tethered to the KICKR via an USB dongle ANT+ connection (Garmin, Kansas City, USA). Data from the KICKR were also recorded at a frequency of 1 Hz. The data from the KICKR was recorded via the Wahoo Utility application run on an electronic tablet (Ipad, Apple, CA, USA) and connected to the unit via Bluetooth. The firmware of the KICKR was updated to the most recent version at the time (v1.3.32).

Protocol

The testing protocol began with a CALRIG-driven warm-up phase of 15 min at 100 W, with the bicycle in a mid-range gear (39x21) and the KICKR set to the default value of 2 in 'Level' mode. In this mode, the resistance against the flywheel increases exponentially as a function of its rotational speed, mimicking outdoor riding.

Following this period, the KICKR was given a 'spindown' – an internal calibration process where the flywheel of the trainer is sped up to 36 km.h-1 and then left to decelerate to 16 km.h-1. During the spindown, the device determines the power required to overcome bearing and belt friction, and sets the zero-offset of strain gauges (http://support.wahoofitness.com/hc/en-us/articles/204281794-How-when-do-I-perform-a-spindown-calibration-; accessed on 08/02/16). The zero offset SRM was also manually reset at this point. After calibration of these units, the CALRIG began an incremental protocol starting at 50 W, which increased 50 W every 2 min, up to and including 400 W. Each unit was tested two non-consecutive times in a random order.

Data analysis

The final 10s of data at each power output of the incremental test were averaged for each device and used to assess the accuracy and reliability of the KICKR and SRM units compared to the CALRIG. This duration was chosen as it was deemed to be a suitable balance between: i) a period long enough to minimize synchronization issues across various devices, and ii) short enough to offer meaningful data resolution. The data from each KICKR's first trial (Trial 1) and the corresponding SRM data were used to assess accuracy, while the data from the second trial (Trial 2) were compared to Trial 1 to assess reliability.

To assess accuracy, the percentage error (device vs. CALRIG) at each power output was determined and subsequently expressed as a mean percentage error (across all power outputs) ± 95% Limits of Agreement (LoA) for each trial (1.96 x SD, in accordance with Martin Bland & Altman, 1986). ± 95% confidence

limits (1.96 x standard error) of the mean error were determined as a measure of inter-unit variability. The coefficient of determination (R^2) for each trial was also calculated. To assess test-retest reliability, the change in percentage error at each power output was calculated for Trial 1 vs. Trial 2. Additionally, the typical error of measurement (TEM), expressed as a coefficient of variation (CV%), was calculated as a measure of trial-to-trial noise (Hopkins, 2015). All data are presented as mean ± 95% confidence limit (CL) unless otherwise stated.

Results

Accuracy

Strong relationships were observed between all KICKR units and the CALRIG (R^2>0.999, P< .001), as well as the SRM vs. CALRIG (R^2= 1.00, P< .001) across 50 to 400 W. The average mean-percentage-error across all power outputs for the five KICKR units compared to CALRIG was -1.5% (ranging between units from -3.1% to 0.1%; ±1.7% 95% CL) in Trial 1 (Figure 2A), with two of the units demonstrating statistically significant differences compared with the CALRIG (Table 1). The average KICKR within-trial 95% LoA was ±3.1% (ranging from ±1.6% to ±4.6%). Comparatively, the mean error of the SRM unit was -1.6% across the range of powers tested (Figure 2B), while the within-trial 95% LoA of the SRM was ±5.4%. However, these estimates appear to be largely skewed by data at 50 W. When the 50 W data was removed, the error was reduced to -0.9% while the 95% LoA was reduced to ±3.4% (Table 1.).

Reliability

The average change in mean-percentage-error across the two trials was 1.3% for all KICKR units (ranging from -0.1% to 4.6%, Table 1). Two KICKR units (1 & 5) demonstrated a statistically significant change in mean error from Trial 1 to Trial 2. The TEM (CV%) however, was homogeneous among units, averaging 1.5% and ranging of between 1.1% and 1.9%. For the SRM, the change in mean percentage error was 0.4%, while the TEM of was 0.7%. (0.6% when not including 50W).

Discussion

The primary finding of this investigation was that based on a convenience sample of 5 units, the KICKR ergometer was typically accurate to within the manufacturer claimed ± 3% across the tested 50 - 400 W range, as measured on a first principles calibration rig. Its -1.5% bias was comparable to the more common and scientifically validated SRM power meter (-1.6% bias); although the SRM did exhibit a high error at 50 W (-7%). Overall, the KICKRs are accurate and typically demonstrate low inter-unit variability, although do not have the same level of reproducibility as the SRM.

At present, there are more power measuring devices available to sports scientists, athletes and coaches than ever before. Of these devices, the SRM is often considered as the 'gold standard' (Gardner et al., 2004; Hopker et al., 2010), with early work validating their use for performance measuring and modelling (Martin, Milliken, Cobb, McFadden, & Coggan, 1998). Studies which have directly compared power meters against a dynamic calibration rig have typically found the SRM to have a lower systematic error that its competitors. In one investigation, the SRM reported a mean error of -0.6% across a 180 - 1320 W range, compared to 1.9% error from a Velotron ergometer (Abbiss et al., 2009). When multiple units of the same device were compared, the SRM also displayed a lower between-unit error (2.3%) when compared to the PowerTap hub (-2.5%) over a range of 50 - 1000 W (Gardner et al., 2004). The results of the present study show that from a convenience sample of 5 units tested twice, the Wahoo KICKR displayed a mean error ranging from -3.1 to 1.5% in a 50 – 400W range. When averaged, the mean error from all units was slightly favourable (-1.5%) when compared to the SRM (-1.6%). If an approximate 2% loss in power through the drivetrain of the bicycle is considered (Gardner et al., 2004), this would further improve the agreement of the KICKR with the CALRIG. However, the mean error for the SRM appears to be largely skewed by the data at 50 W. The

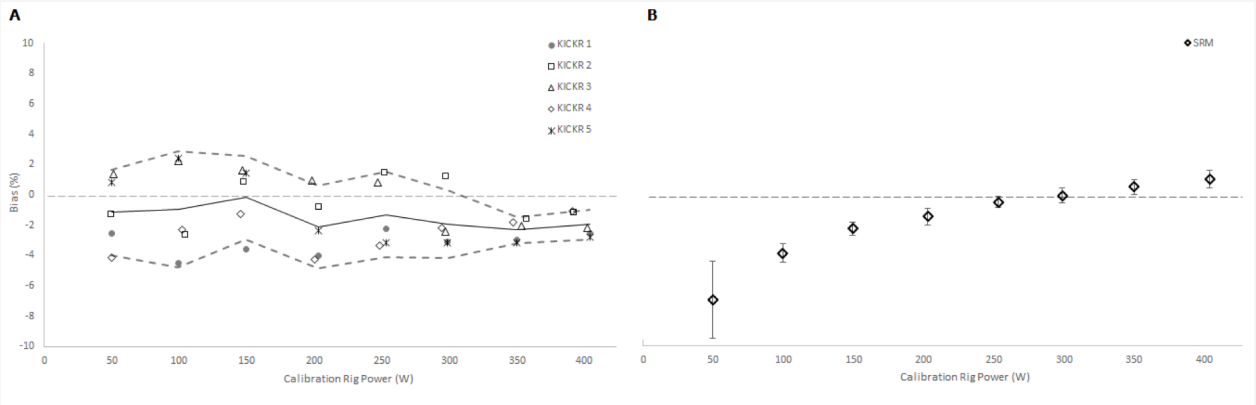

Figure 2. Modified Bland-Altman plot Bland-Altman plot of the A) percentage error (vs. CALRIG) for all five KICKR units at each power output, as well as the average percentage error at each power output (solid line) and between-unit ±95% CL (dotted lines) between the KICKRs and CALRIG for Trial 1; and B) the mean error and between-trial ±95% CL (error bars) for the SRM compared to CALRIG for the same Trials.

Table 1. Comparison of the mean error for all power meters tested in comparison to the calibration rig in the range of 50-400 W, measured in 50 W increments.

	Trial 1		Trial 2		Reliability	
	Mean Error (%)	*±95% LoA*	*Mean Error (%)*	*±95% LoA*	*Difference in Mean (Trial 2 – 1)*	*TEM (CV%)*
KICKRs						
KICKR 1	-3.1*	1.6	1.5*	2.3	4.6**	1.4
KICKR 2	-0.5	2.9	-1.1	4.0	-0.7	1.9
KICKR 3	0.1	3.8	-0.1	2.8	-0.1	1.8
KICKR 4	-1.2	4.6	-0.8	3.5	0.5	1.3
KICKR 5	-2.5*	2.4	-0.2	2.8	2.4**	1.1
Mean	-1.5	3.1	-0.1	3.1	1.3	1.5
±95% CL *(inter-unit variability)*	1.7		1.2		2.7	0.4
SRM *(n=5)*						
Mean	-1.6	5.4	-1.3	5.2	0.4	0.7
SRM (≥100 W)						
Mean	-0.9	3.4	-0.5	3.2	0.4	0.6

* Significantly different vs. CALRIG; ** Significant change between Trial 1 & 2; CL: Confidence limits; LoA: Limits of Agreement; TEM: Typical Error Of Measurement, expressed as a coefficient of variation (CV%)

significance of accurate data in this power range is arguably less important for trained individuals, and if the outlying data at 50 W is removed, the mean error of the SRM is greatly reduced to -0.9%. A greater error in this low power range has also been reported in the only other available study to assess the SRM at this level (Gardner et al., 2004), suggesting it may be a common systematic error for the SRM. Regardless, the data collected in the present study confirms the revered accuracy that has been associated with the SRM, and suggests the KICKR is capable of a similar standard. The validity of both devices is further supported by all units demonstrating a strong relationship (R^2>0.999) with the CALRIG.

The only other paper, to date, that has assessed the accuracy of the KICKR reported a -1.1% bias across a 250 – 700 W range (Zadow, Kitic, Wu, Smith, & Fell, 2016), similar to the mean -1.5% bias across 50 - 400 W we have reported in the current investigation. However, Zadow and colleagues found a much higher error (4.5%) in the lower power range of 100 – 200 W; with the present investigation finding the bias to not greatly fluctuate between high and low ranges (Figure 1). The reason for this discrepancy is unclear, although the reported 4.5% error may be partially inflated due to one atypical outlying measure (>2.5 SD from mean). It should be noted that Zadow and colleagues used only one KICKR. From our data of five KICKRs, it does not appear that this characteristic of greater errors <200 W is a systematic error across all devices.

Of equal importance to the accuracy of power measuring devices, is the reliability of a unit such that accurate assessment of changes in performance can be made (Paton & Hopkins, 2001). Nonetheless, this aspect is not commonly examined in power meter investigations. Key to quantifying reliability is assessing the TEM (i.e. the within-unit random noise) of a device and the change in the mean (% bias in this instance) between trials (Hopkins, 2000). Here, we found the SRM to be a highly reliable device, with minimal change in mean bias from Trials 1 & 2, suggesting a low systematic error (i.e. its accuracy was maintained across Trials). This was supported by a relatively low TEM (CV 0.7%), indicating a low level of random error and thus, a high degree of test-retest reproducibility. Comparatively, each KICKR unit displayed a higher degree of variability between Trials, with a majority displaying a greater shift in the bias, and a larger TEM (CV 1.1 – 1.9%). Two of the units (1 & 5) in particular experienced significant shifts in the mean bias. The source of this deviation is unclear, with possible sources of error from the warm-up and calibration phase, and the bicycle-trainer-CALRIG setup despite our best efforts for consistency in methodology. This carries implications for monitoring changes in performance, with the smallest, worthwhile change that may be ascertained from a test suggested to be no less than the TEM (Paton & Hopkins, 2001). Regardless, the TEM demonstrated by the KICKR in this investigation compares favourably to other investigated power meters, including the Wattbike (CV 2.6%; Hopker et al., 2010), PowerTap (CV 1.8%; Bertucci, Duc, Villerius, Pernin, et al., 2005), Axiom Powertrain (CV 2.2%; Bertucci, Duc, Villerius, & Grappe, 2005) and Ergomo Pro (CV 2.3 – 4.1%; Duc, Villerius, Bertucci, & Grappe, 2007; Kirkland, Coleman, Wiles, & Hopker, 2008); which were all deemed to be a low and acceptable level of error. In this regard, the KICKR appears to be a reliable power meter when compared against a dynamic CALRIG, although its suitability for assessing very small changes in human performance requires further investigation.

Despite generally displaying a relatively high level of accuracy and reliability, it is apparent from our sample of 5 KICKR units that devices will differ. In our study, KICKR 1 showed the highest degree of inaccuracy in Trials 1 and 2 respectively, as well as the greatest change in bias (Δ 4.5%) across trials. Although we have shown the KICKR may be as accurate as the SRM, it is important for users to assess each individual unit. Even so, one of the major disadvantages of the KICKR is the inability to adjust the slope of the calibration curve without proprietary equipment. Even with this equipment, only a two-point calibration is possible. The SRM on the other hand, may be manually adjusted to match either a multi-point static calibration

(as performed prior to testing in this study), or a more robust dynamic calibration (e.g. using the data collected against the CALRIG to form a regression equation). Though the KICKR has an internal calibration process (i.e. 'spindown'), the more ecologically valid dynamic calibration (Hopkins et al., 2001) is not currently possible.

Practical application

The Wahoo KICKR is a tool which allows the use of most bicycles as ergometers capable of performance assessment in a cost-effective manner. This is likely to see the KICKR adopted by more athletes, coaches and laboratories – therefore it is important to establish the accuracy and reliability of this device. Here, we found the accuracy can be comparable to that of a statically calibrated SRM across a 50-400 W range, although inter-unit variability means this is not always the case. Although not as reliable as the SRM, the KICKR also demonstrated an acceptable level of error. Now that the KICKR's performance against a dynamic calibration rig has been established and validated, research should now focus on its applicability in monitoring human performances.

Study Limitations

One of the limitations to the findings of our study is the range of power examined on the KICKR. Although the range of power tested in the present study would cover a majority of human performances in a graded exercise test, we were unable to test the accuracy of the power meter above 400 W due to limitations in the calibration rig's capacity. While we are unable to confirm the accuracy of the KICKR beyond this point and up to the advertised maximum power of 1550 W, a recently published study has examined the accuracy of the KICKR up to 1000 W (Zadow et al., 2016).

Additionally, it is important to note that the data was analysed in 10 s averages in our investigation. Care should be taken by users if trying to assess time periods less than 10 s (for e.g. sprinting efforts), particularly given that flywheel based power meters have previously been reported to have a hysteresis (Abbiss et al., 2009). This lag between input and output is more likely to impact analysis of shorter time periods than longer ones, depending on the extent of the delay. This consideration is especially pertinent to more stochastic, variable exercise. Further work is required to establish if KICKRs are still suitable in more 'ecological' situations were fluctuations in power are more common, including performances in human subjects. This should include focus on prospective effects of pedaling cadence, and the potential drift of power over longer efforts, particularly as similar devices have been reported to drift over time (Bertucci, Duc, Villerius, & Grappe, 2005).

Conflict of interest

The authors confirm that there are no conflicts of interest in this article.

References

1. Abbiss, C. R., Quod, M. J., Levin, G., Martin, D. T., & Laursen, P. B. (2009). Accuracy of the Velotron ergometer and SRM power meter. Int J Sports Med, 30(2), 107-112. doi: 10.1055/s-0028-1103285
2. Bertucci, W. (2012). Analysis of the agreement between the Fortius cycling ergometer and the PowerTap powermeter PO during time trials of 6 and 30 min. Computer Methods in Biomechanics and Biomedical Engineering, 15(sup1), 212-214. doi: 10.1080/10255842.2012.713604
3. Bertucci, W., Duc, S., Villerius, V., & Grappe, F. (2005). Validity and Reliability of the Axiom Powertrain Cycle Ergometer When Compared with an SRM Powermeter. Int J Sports Med, 26(01), 59-65. doi: 10.1055/s-2004-817855
4. Bertucci, W., Duc, S., Villerius, V., Pernin, J. N., & Grappe, F. (2005). Validity and reliability of the PowerTap mobile cycling powermeter when compared with the SRM Device. Int J Sports Med, 26(10), 868-873. doi: 10.1055/s-2005-837463
5. Duc, S., Villerius, V., Bertucci, W., & Grappe, F. (2007). Validity and reproducibility of the ErgomoPro power meter compared with the SRM and Powertap power meters. Int J Sports Physiol Perform, 2(3), 270-281.
6. Gardner, A. S., Stephens, S., Martin, D. T., Lawton, E., Lee, H., & Jenkins, D. (2004). Accuracy of SRM and power tap power monitoring systems for bicycling. Med Sci Sports Exerc, 36(7), 1252-1258.
7. Hopker, J., Myers, S., Jobson, S. A., Bruce, W., & Passfield, L. (2010). Validity and reliability of the Wattbike cycle ergometer. Int J Sports Med, 31(10), 731-736. doi: 10.1055/s-0030-1261968
8. Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. Sports Med, 30(1), 1-15.
9. Hopkins, W.G. (2015). Spreadsheets for Analysis of Validity and Reliability. Sportscience, 19, 36-42
10. Hopkins, W. G., Hawley, J. A., & Burke, L. M. (1999). Design and analysis of research on sport performance enhancement. Med Sci Sports Exerc, 31(3), 472-485.
11. Hopkins, W. G., Schabort, E. J., & Hawley, J. A. (2001). Reliability of Power in Physical Performance Tests. Sports Medicine, 31(3), 211-234. doi: 10.2165/00007256-200131030-00005
12. Kirkland, A., Coleman, D., Wiles, J. D., & Hopker, J. (2008). Validity and Reliability of the Ergomo®pro Powermeter. Int J Sports Med, 29(11), 913-916. doi: 10.1055/s-2008-1038621
13. Martin Bland, J., & Altman, D. (1986). Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. The Lancet, 327(8476), 307-310. doi: 10.1016/S0140-6736(86)90837-8
14. Martin, J. C., Milliken, D. L., Cobb, J. E., McFadden, K. L., & Coggan, A. R. (1998). Validation of a mathematical model for road cycling power. Journal of applied biomechanics, 14, 276-291.
15. Novak, A. R., Stevens, C. J., & Dascombe, B. J. (2015). Agreement between LeMond Revolution cycle ergometer and SRM power meter during power profile and ramp protocol assessments (Vol. 4).
16. Paton, C. D., & Hopkins, W. G. (2001). Tests of cycling performance. Sports Med, 31(7), 489-496.
17. Wooles, A., Robinson, A., & Keen, P. (2005). A static method for obtaining a calibration factor for SRM bicycle power cranks. Sports Engineering, 8(3), 137-144. doi: 10.1007/BF02844014
18. Zadow, E. K., Kitic, C. M., Wu, S. S., Smith, S. T., & Fell, J. W. (2016). Validity of Power Settings of the Wahoo KICKR Power Trainer. Int J Sports Physiol Perform. doi: 10.1123/ijspp.2015-0733