***Title page (including all author details and affiliations)**

Critical power: How different protocols and models affect its determination

Felipe Mattioni Maturana[a], Federico Y. Fontana[b], Silvia Pogliaghi[b], Louis Passfield[c], Juan M. Murias[a]

[a] Faculty of Kinesiology, University of Calgary, Calgary, AB, Canada; [b] Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Verona, Italy; [c] School of Sport & Exercise Sciences, University of Kent, Chatham Maritime, Kent, United Kingdom

Corresponding author: Juan M. Murias

Faculty of Kinesiology, University of Calgary

KNB 434, 2500 University Dr NW

Calgary, AB, Canada, T2N 1N4

e-mail: jmmurias@ucalgary.ca

tel. +1 403 220 7955, fax. +1 403 220 0448

Word count: 2898

Abstract count: 246

Tables: 2

Figures: 1

Critical power: How different protocols and models affect its determination

Abstract

In cycling, critical power (CP) and work above CP (W') can be estimated through linear and nonlinear models. Despite the concept of CP representing the upper boundary of sustainable exercise, overestimations may be made as the models possess inherent limitations and the protocol design is not always appropriate. **Objectives:** to measure and compare CP and W' through the exponential (CP_{exp}), 3-parameter hyperbolic ($CP_{3\text{-}hyp}$), 2-parameter hyperbolic ($CP_{2\text{-}hyp}$), linear (CP_{linear}), and linear 1/time ($CP_{1/time}$) models, using different combinations of TTE trials of different durations (approximately 1 to 20 min). **Design:** repeated measures. **Methods:** Thirteen healthy young cyclists (26±3yrs; 69.0±9.2kg; 174±10cm; 60.4±5.9mL·kg^{-1}·min^{-1}) performed five TTE trials on separate days. CP and W' were modeled using two, three, four, and/or five trials. All models were compared against a criterion method ($CP_{3\text{-}hyp}$ with five trials; confirmed using the leaving-one-out cross-validation analysis) using smallest worthwhile change (SWC) and concordance correlation coefficient (CCC) analyses. **Results:** CP was considerably overestimated when only trials lasting less than 10 min were included, independent of the mathematical model used. Following CCC analysis, a number of alternative methods were able to predict our criterion method with almost a perfect agreement. However, the application of other common approaches resulted in an overestimation of CP and underestimation of W', typically these methods only included TTE trials lasting less than 12 min. **Conclusions:** Estimations from $CP_{3\text{-}hyp}$ were found to be the most accurate, independently of TTE range. Models that include two trials between 12 and 20 min provide good agreement with the criterion method (for both CP and W').

Key-Words: power-time relationship; time-to-exhaustion; linear model; nonlinear model; exercise intensity domains.

Introduction

Since first introduced by Monod and Scherrer[1] as the maximal capacity of a muscle, or muscle group to perform work for a prolonged period of time, the concept of critical power (CP) has been widely used as it presents a useful approximation of the endurance capacity of an individual[2-4]. Typically CP is determined from a series of 5 time-to-exhaustion trials (TTE) conducted at severe exercise intensities[5-7]. However, several studies suggest that estimates of CP can vary and are influenced by the test protocol design. Factors such as the particular model used, and the duration of the TTE trials can change the CP calculated from the model[8-11]. Researchers use varying models to estimate CP, which are derived from a range of two to seven TTE trials that are not standardized in terms of their duration. Although we note that the most commonly used method is probably one employing four to five trials and fitted with the two-parameter hyperbolic model ($CP_{2\text{-}hyp}$)[7].

Different studies have focused on either the influence of changing the mathematical model, or the number of repetitions on the derived value for CP. For example, Gaesser et al.[8], Bull et al. [10], and Bergstrom et al. [11] investigated the influence of different mathematical models such as exponential (CP_{exp}), three parameter hyperbolic ($CP_{3\text{-}hyp}$), $CP_{2\text{-}hyp}$, linear (CP_{linear}), and linear 1/time ($CP_{1/time}$) on the determination of CP and the work above CP (W'). These three studies found $CP_{3\text{-}hyp}$ and CP_{exp} result in different estimations of CP. Bishop et al. [9] asked their participants to perform five TTE trials ranging from 1 to 10 minutes in duration in order to evaluate the influence of the length of TTE trials on CP parameter predictions. Using data from only three of the five trials CP was modelled with CP_{linear} and $CP_{2\text{-}hyp}$. Bishop et al. found that a significant difference in modelled CP when the three shortest trials (i.e., $CP_{1,2,3}$), the three longest trials (i.e., $CP_{3,4,5}$), or the first, the third, and the fifth trials (i.e., $CP_{1,3,5}$) were selected. Consequently, the authors suggested that TTE trials of widely varying duration should be used to minimize the influence of shorter trials when modelling CP. However, this investigation did not fit the data from all five TTE trials, and was also limited modelling CP using TTE rides of less than 10 min, about half the longest duration recommended by Morton [7]. Moreover, the aforementioned studies lacked comparisons of the effects of using different mathematical methods and range of TTE trials on W' outcomes.

Given the variety of approaches used in the literature and the effects of different models and combinations of TTE trials, and the lack of a complete comparison of estimations of CP and W', the present investigation aimed to examine the effect of number and range of TTE trials, and equation model on CP and W'. Specifically we modelled CP and W' using combinations of two to five TTE trials with a variety of different mathematical approaches (CP_{exp}, $CP_{3\text{-}hyp}$, $CP_{2\text{-}hyp}$, CP_{linear}, and $CP_{1/time}$).

Methods

Thirteen healthy young participants (9 men and 4 women; mean ± SD values: age, 26 ± 3 yr; body mass, 69.0 ± 9.2 kg; height, 174 ± 10 cm) volunteered and gave written informed consent to participate in the study. All participants had previous recreational or competitive cycling experience at the provincial level. Participants were nonsmokers, with no musculoskeletal or cardiorespiratory conditions. The full testing protocol was completed in 3 ± 1 weeks and consisted of: i) a preliminary maximal ramp incremental test for determination of maximal $\dot{V}O_2$ ($\dot{V}O_{2max}$), and peak power output (PO_{peak}); and ii) five TTE trials for estimation of CP. All procedures were conducted in an environmentally controlled laboratory (i.e. temperature ~21°C, relative humidity ~36%), at a similar time of the day for each participant, with each test performed on separate days, with a minimum interval of 24 h and a maximum interval of 72 h (most typically 48 h) between tests to ensure appropriate recovery between trials. Participants were instructed to keep their water and carbohydrate intake consistent throughout the protocol, and they were requested not to engage in vigorous physical activity for 24 h prior to each test. Participants were asked not to consume caffeine less than 12 h prior to the test. This study was approved by the Conjoint Health Research Ethics Board of the University of Calgary. The results from $CP_{2\text{-}hyp}$ using five TTE trials have been published as part of a separate study comparing CP with the maximal lactate steady-state [5].

All exercise tests were performed on an electromagnetically braked cycle ergometer (Velotron Dynafit Pro, Racer Mate, Seattle, WA, USA). Breath-by-breath pulmonary gas exchange, ventilation and heart rate (HR) were continuously measured using a metabolic cart (Quark CPET, COSMED, Rome, Italy), as previously described [12]. Calibration was performed before each test as recommended by the manufacturer. Breath-by-breath $\dot{V}O_2$ data were edited as follows: data points that were 3 SD from the local mean were considered outliers and then removed [13]; trials were time-aligned to the onset of exercise (i.e. time zero representing the onset of the ramp incremental exercise), and averaged into 30-s time bins. $\dot{V}O_{2max}$ was considered as the highest 30-s $\dot{V}O_2$ average throughout the ramp incremental test. PO_{peak} was established as the highest power output achieved at the end of the ramp incremental test.

For the ramp incremental test, the baseline consisted of participants cycling at 50 W for 4 min, as suggested by Boone and Bourgois [14], followed by either 1 W every 2 s (30 $W{\cdot}min^{-1}$) (men) or 1 W every 2.4 s (25 $W{\cdot}min^{-1}$) (women) increase in PO.

For the estimation of CP, each participant performed five constant-power output trials to exhaustion which ranged from approximately 1 – 20 min, as recommended by Morton [7]. The first three TTE trials were performed at 80, 95 and 110% of PO_{peak} (as determined from the preliminary ramp incremental test). The order of the tests was randomly assigned. Subsequently, the other two power outputs were determined to generate an even distribution of TTE between the five trials. Each test was preceded by a 4-min baseline at 20 W, followed by a square-wave transition to the predetermined PO.

For all TTE trials, participants cycled at their preferred pedal cadence (range, 70-105 rpm), which was determined during the preliminary ramp incremental test. The moment of exhaustion was deemed to occur when participants failed to maintain the cadence within 5 rpm of their preferred rate for more than 5 s despite strong verbal encouragement. Participants were blinded to the elapsed time, but they received visual feedback on their pedal cadence.

CP was modelled as follows:

i. $CP_{exp} \rightarrow PO = CP + (P_{max} - CP) * exp(-t / \tau)$ *Hopkins et al.* [15]

ii. $CP_{3\text{-}hyp} \rightarrow t = (W' / PO - CP) + (W' / CP - P_{max})$ *Morton* [16]

iii. $CP_{2\text{-}hyp} \rightarrow t = W' / (PO - CP)$ *Hill* [17]

iv. $CP_{linear} \rightarrow W_{lim} = W' + CP * t$ *Moritani et al.* [18]

v. $CP_{1/time} \rightarrow PO = W' * (1 / t) + CP$ *Whipp et al.* [19]

where P_{max} is the maximal instantaneous power (in watts), τ an undefined time constant, and W_{lim} is the work done (i.e., PO * t) in each predictive trial (in Joules).

When the model was fitted using four trials, two combinations were used: trials *1 to 4*, and trials *2 to 5*. Using three trials, four combinations were performed: trials *1, 2, 3*; trials *1, 3, 5*; trials *2, 3, 4*; and trials *3, 4, 5*. Finally, when using two trials in the linear models, four combinations were tested: trials *1 and 2*; trials *1 and 5*; trials *3 and 4*; and trials *4 and 5*. Importantly, not every possible combination was reported to avoid superfluous comparisons that would not add predicting value to the model. Instead, we selected the combination of methods that would result in a wide combination of TTE, as well as those often used in the literature. See Table 1 for details on the exercise intensities and durations of the aforementioned TTE trials.

All data editing, processing, and modeling were performed using OriginLab version 9.2 (OriginLab, Northampton, MA).

Data are presented as means ± SD. 90 % confidence intervals were calculated and used as a measure of uncertainty (the likely limit of the true value in the population [20]) around each CP and W' values derived from the different methods proposed. Differences between methods were quantified by calculating chances that the true value of a difference was substantial or greater than the smallest worthwhile change (see below). To perform these calculations, we assumed that a substantial difference (in either direction, positive or negative) was larger than 8 W (3.2 %) and 1500 J (6.5%) (these are calculated as a constant factor (0.2) multiplied by the between-subjects standard deviation [20] around the criterion-method average

CP and W' values of the 3-parameter hyperbolic method using the all trials (i.e., $CP_{3\text{-}hyp(1,2,3,4,5)}$) as described below. The above calculated thresholds were defined as the smallest worthwhile changes perceived to be practically meaningful for both CP and W'. Thresholds for assigning qualitative terms to chances of substantial effects were as follows: <1 %, *almost certainly not*; 1-5 %, *very unlikely*; 5-25 %, *unlikely* or *probably not*; 25-50 %, *possibly not*; 50-75 %, *possibly*; 75-95 %, *likely* or *probably*; 95-99 % *very likely*; >99 %, *almost certain*[20,21]. Here the criterion value chosen to declare an effect as *likely/possibly* vs *unclear* is based on a probabilistic approach. In Figure 1 and Figure 2, the exact probability of the difference is reported. Effect sizes of each difference (Cohen's d, ranked as *trivial* (0-0.19), *small* (0.20-0.49), *medium* (0.50-0.79) and *large* (0.80 and grater) [22] are also reported as objective and standardized measures of magnitude of effects and as alternative meaningfulness metrics [23]. In effect size calculation, the SD of $CP_{3\text{-}hyp(1,2,3,4,5)}$ was used to standardize the mean difference for each contrast [24]. The appropriateness of the $CP_{3\text{-}hyp(1,2,3,4,5)}$ model as "criterion method" for our data was determined by testing how well this model fitted the observed data. CP parameter estimates of each method as well as the ability of each model to generalize to new data were tested using the leave-one-out cross-validation (LOOCV) approach. The model that fits the data most closely for both CP and W' was confirmed as $CP_{3\text{-}hyp(1,2,3,4,5)}$.

The measurement agreement between the criterion method and each other models or number of trials was assessed by evaluating Lin's concordance correlation coefficient (CCC)[25]. Concisely, this metric indicates the degree to which the relationship between two variables approximates the perfect agreement (*i.e.* line-of-identity)[26]. The CCC was interpreted using the following criterion ranges: almost perfect agreement (CCC > 0.99), substantial agreement (0.95 > CCC < 0.99), moderate agreement (0.90 < CCC > 0.95), and poor agreement (CCC < 0.90)[27]. Additionally, the RMSE and the slope/intercept resulting from the above regression analyses were used to *i)* indicate the typical error that may be expected when using any "inadequate" model to directly estimate the criterion model (*i.e.* $CP_{3\text{-}hyp(1,2,3,4,5)}$) and *ii)* to understand whether this error was better or worse at high or low values of CP (see Supplementary Material 1 for the results of the above mentioned analysis).

The statistical analysis was performed using STATA (Version 14, Texas, USA) and α was set in advance at the 0.05 level; statistical significance was accepted when $p < \alpha$.

Results

Group mean absolute and relative $\dot{V}O_{2max}$ were 4.17 ± 0.68 $L\cdot min^{-1}$ (range: 2.85 – 5.08 $L\cdot min^{-1}$) and 60.4 ± 5.9 $mL\cdot kg^{-1}\cdot min^{-1}$ (range: 50.7 – 68.1 $mL\cdot kg^{-1}\cdot min^{-1}$), respectively. Group mean PO_{peak} was 376 ± 54 W (range: 274 – 448 W).

Group mean duration, corresponding exercise intensities of TTE trials, and mechanical work (i.e., *time* * *PO*) for CP estimations are summarized in Table 1. The mean duration of trials ranged from 1.7 to 19.4 min. Group mean parameter estimates (i.e., CP and W') from the combinations performed between number of trials *vs* mathematical models are displayed in Table 2.

Based on the LOOCV analysis, the model that predicts data most accurately was confirmed as the $CP_{3\text{-}hyp(1,2,3,4,5)}$ (R^2 = 0.99, 95% CI [248 255], RMSE = 26.5 W). Figures 1 and 2 show the mean difference between each model and the criterion model for CP and W', respectively. The difference for the majority of the alternative methods was declared *unclear* for CP. However, for the methods $CP_{2\text{-}hyp(1,2,3)}$, $CP_{linear(1,2)}$, $CP_{linear(1,2,3)}$, $CP_{1/time(1,2,3)}$, $CP_{1/time(1,2,3,4)}$, and $CP_{1/time(1,2,3,4,5)}$ the difference in relation to the criterion method was considered *likely positive* (i.e., overestimation). When using CP_{exp}, the CP estimates were consistently higher than those observed in the criterion method, with the chance of an overestimation declared *very likely positive* (Figure 1).

The difference in W' among almost all the alternative methods (20 out of 31) was considered *likely negative* (i.e., underestimation) (Figure 2). With respect to the criterion method ($CP_{3\text{-}hyp(1,2,3,4,5)}$) a number of alternative methods resulted in a very small chance of underestimating W': $CP_{3\text{-hyp}(1,2,3,4)}$, $CP_{3\text{-hyp}(2,3,4,5)}$, $CP_{2\text{-hyp}(3,4,5)}$, $CP_{2\text{-hyp}(2,3,4,5)}$, $CP_{2\text{-hyp}(1,2,3,4,5)}$, as well as CP_{linear} and $CP_{1/time}$ using the trials (3,4), (4,5), and (3,4,5). In Figure 2 it is notable that the inclusion of trials lasting less than 10 min (i.e., trials 1 – 3) caused a substantial underestimation of W', whereas the inclusion of trial 5 (approximately 20 min, on average) led to the best approximation to the criterion method.

The results of the CCC analysis are presented in Supplementary Material 1.

Discussion

The main findings of this study were that: i) using TTE trials lasting less than 10 min (i.e., trials 1-3) to model CP resulted in consistently higher values than those using the criterion method, and a considerable underestimation of W'; ii) when longer TTE trials were included in the model (between approximately 12 to 20 min), the estimations of CP were similar to those observed for the criterion method; and iii) $CP_{1/time}$ may provide an accurate CP and W' estimation, as long as TTE trials lasting less than 7 min are not included in the mathematical model.

Given the popularity of CP as a measure of sustainable exercise intensity, accurate determination of CP is important. Since its introduction more than 60 years ago [1], different protocol designs have been used to determine CP. For example, whereas Poole [28] stated that predictive trials should range between 1 and 10 min, Morton [7] suggested that longer trials, ranging from approximately 1 to 20 min, should be included in order to model a power output that more realistically predicts the upper boundary of sustainable endurance exercise.

Housh et al. [29] studied the effect of different combinations of TTE trials when modelling CP. The authors compared CP modelled using two and three TTE trials against a pre-determined criterion method (four trials), in an attempt to find the optimal protocol. These researchers found that when using the shortest (~1 min) and the longest (~10 min) trials, the estimation was the most accurate and presented the lowest standard error of the estimate. Therefore, they suggested that CP could be accurately estimated using two trials, lasting 1 and 10 min. Bishop et al. [9] conducted five TTE trials and estimated CP using different combinations of three trials. The authors found significant differences in the CP values when using any combination of three TTE trials, across a range of 1 to 10 min. The present study shows that using this range of predictive trials (i.e., 1 to 10 min) the modelled CP is consistently higher than that obtained using the criterion method. Furthermore, this effect is independent of the mathematical model used. In fact, such differences in relation to the criterion method were always substantially higher than our minimum detectable difference of 8 W. This finding implies that a time range within 1 to 10 min will likely result in an overestimation of CP irrespective of the model selected (Figure 1). Additionally, CCC analysis (Supplementary Material 1) also revealed that models including only shorter trials (*i.e.* less than 10 min) resulted in: *i)* poor agreement with the criterion method (CCC < 0.90) and *ii)* a disproportionally higher or lower estimation of CP (as indicated by the reported slope and intercept values). This may also have implications for the interpretation of previous studies where the CP has been modelled only with TTE trials lasting less than 10 min.

Based on the results from this and previous investigations, the mathematical model can have a significant impact on the predicted CP [8,30]. In other words, for a given set of data points (e.g., three TTE trials), different CP predictions may be generated by each of the models. Our results show that several models allow the predicted CP value to be overestimated (with the probability ranging from 75 to 99%, see Figure 1). In contrast, the 3-parameter hyperbolic model (i.e., $CP_{3\text{-hyp}}$) appears to provide the most accurate approach, regardless of the TTE trials modelled. This supports the contention that $CP_{3\text{-hyp}}$ overcomes the limitation of other linear and nonlinear models that assume an infinite power as *time* approaches zero[16]. This limitation is addressed by adding a third-parameter to the CP model, the so-called maximal instantaneous power (P_{max}). Interestingly, $CP_{2\text{-hyp}}$ and CP_{linear} only overestimated CP, when the range of predictive trials all had a duration of less than 10 min. This suggests that accurate predictions of CP can be achieved provided longer TTE trials are included. Lastly, and in contrast with previous investigations [8,11], $CP_{1/time}$ may provide an accurate measure of CP if trials longer than 10 min are included in the model (e.g., $CP_{1/time(4,5)}$ and $CP_{1/time(1,5)}$). Importantly, it should be noted that CP may not always reflect the highest boundary of physiological steady-state with prolonged exercise, as shown in previous investigations[5,31]. However, by using the most appropriate testing method (i.e., model and range of TTE trials), it is likely that the CP value will more closely approximate the highest PO associated with a metabolic steady-state, and will provide better estimations of TTE for any given intensity above CP for performance prediction.

The accurate prediction of CP is possible with a range of different CP models provided that longer TTE trials are included. This means that an accurate prediction of CP is possible using simpler mathematical models and fewer tests (see also Supplementary Material 1). In this context, based on the present results, CP may be estimated using either the CP_{linear} or $CP_{1/time}$ models with as few as two predictive trials if they range from approximately 7 to 20 min (e.g., $CP_{linear(3,4)}$, $CP_{1/time(4,5)}$). These data are most relevant in "field conditions" where time-efficiency (i.e. reducing the number of repetitions to minimize the time commitment of athletes to testing) is a priority and where testing results can be combined with perception of effort[32] towards the fine-tuning of training intensity. However, when maximal accuracy and repeatability are required, such as in a longitudinal research design, researchers should use several TTE trials and a model that possesses high accuracy (i.e., hyperbolic) for CP estimation.

Alongside CP, the accuracy of W’ is important for performance, as it delineates exercise capacity in the severe-intensity domain[33]. As CP models are often used for predicting the optimal time for a given distance, a reliable measure of W’ becomes crucial for the success of coaches and sport scientists in the final outcome of a race. As observed in the present results, inclusion of TTE trials lasting less than 10 min results in an underestimation of W’, whereas the inclusion of two TTE trials ranging between 12 and 20 min in the model yielded the most accurate W’ estimations when compared to our criterion method.

In conclusion, estimations from $CP_{3\text{-}hyp}$ provided the most accurate and generalizable approach for CP and W’ calculation (i.e., the model that was the least affected by protocol design). Accurate estimations of CP can be made with models that use fewer exercise tests and simpler analyses, such as $CP_{2\text{-}hyp}$, CP_{linear}, and

$CP_{1/time}$. However, for these methods to express their accuracy, longer TTE trials ranging from approximately 7 to 20 min should be included in the model. Modeling only TTE trials lasting less than 10 min may lead to a considerable overestimation of CP especially when using CP_{linear} and $CP_{1/time}$, as well as underestimation of W'.

Practical Implications

- The use of only TTE trials lasting less than 12 min not only overestimates CP, but also underestimates W'. Scientists and coaches should ensure that two TTE trials ranging between 12 and 20 min are performed to provide more accurate CP and W' estimations;
- CP_{linear} and $CP_{1/time}$ models using two TTE trials lasting between 7 and 20 min may provide accurate estimations of CP and, combined with self-perception of effort, can offer a valid option when time-efficient models are preferable;
- $CP_{3\text{-}hyp}$ is not affected by protocol design and $CP_{2\text{-}hyp}$ only overestimates CP when only trials lasting less than 10 min are used. This suggests that these models might be more appropriate for research purposes where reducing measurement variability is critical.

References

1. Monod H, Scherrer J. The Work Capacity of a Synergic Muscular Group. *Ergonomics*. 1965; 8(1-4):329-338.
2. Keir DA, Fontana FY, Robertson TC, et al. Exercise Intensity Thresholds: Identifying the Boundaries of Sustainable Performance. *Med Sci Sports Exerc*. 2015; 47(9):1932-1940.
3. Poole DC, Burnley M, Vanhatalo A, et al. Critical Power: An Important Fatigue Threshold in Exercise Physiology. *Med Sci Sports Exerc*. 2016.
4. Poole DC, Ward SA, Gardner GW, et al. Metabolic and respiratory profile of the upper limit for prolonged exercise in man. *Ergonomics.* 1988; 31(9):1265-1279.
5. Mattioni Maturana F, Keir DA, McLay KM, et al. Can measures of critical power precisely estimate the maximal metabolic steady-state? *Appl Physiol Nutr Metab*. 2016; 41(11):1197-1203.
6. Vanhatalo A, Doust JH, Burnley M. Determination of critical power using a 3-min all-out cycling test. *Med Sci Sports Exerc*. 2007; 39(3):548-555.
7. Morton RH. The critical power and related whole-body bioenergetic models. *Eur J Appl Physiol*. 2006; 96(4):339-354.
8. Gaesser GA, Carnevale TJ, Garfinkel A, et al. Estimation of critical power with nonlinear and linear models. *Med Sci Sports Exerc*. 1995; 27(10):1430-1438.
9. Bishop D, Jenkins DG, Howard A. The critical power function is dependent on the duration of the predictive exercise tests chosen. *Int J Sports Med*. 1998; 19(2):125-129.
10. Bull AJ, Housh TJ, Johnson GO, et al. Effect of mathematical modeling on the estimation of critical power. *Med Sci Sports Exerc*. 2000; 32(2):526-530.
11. Bergstrom HC, Housh TJ, Zuniga JM, et al. Differences among estimates of critical power and anaerobic work capacity derived from five mathematical models and the three-minute all-out test. *J Strength Cond Res*. 2014; 28(3):592-600.

12. De Roia G, Pogliaghi S, Adami A, et al. Effects of priming exercise on the speed of adjustment of muscle oxidative metabolism at the onset of moderate-intensity step transitions in older adults. *Am J Physiol Regul Integr Comp Physiol.* 2012; 302(10):R1158-1166.
13. Lamarra N, Whipp BJ, Ward SA, et al. Effect of interbreath fluctuations on characterizing exercise gas exchange kinetics. *J Appl Physiol.* 1987; 62(5):2003-2012.
14. Boone J, Bourgois J. The oxygen uptake response to incremental ramp exercise: methodogical and physiological issues. *Sports Med.* 2012; 42(6):511-526.
15. Hopkins W, Edmond I, Hamilton B, et al. Relation between power and endurance for treadmill running of short duration. *Ergonomics.* 1989; 32(12):1565-1571.
16. Morton RH. A 3-parameter critical power model. *Ergonomics.* 1996; 39(4):611-619.
17. Hill DW. The critical power concept. A review. *Sports Med.* 1993; 16(4):237-254.
18. Moritani T, Nagata A, deVries HA, et al. Critical power as a measure of physical work capacity and anaerobic threshold. *Ergonomics.* 1981; 24(5):339-350.
19. Whipp B, Huntsman D, Storer T, et al. A constant which determines the duration of tolerance to high-intensity work. *Federation Proceedings.* Vol 41: FEDERATION AMER SOC EXP BIOL 9650 ROCKVILLE PIKE, BETHESDA, MD 20814-3998; 1982:1591-1591.
20. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Sportscience.* 2005; 9:6-14.
21. Hopkins WG. A scale of magnitudes for effect statistics. *A new view of statistics.* 2002; 502.
22. Cumming G. The new statistics: Why and how. *Psychol. Sci.* 2014; 25(1):7-29.
23. Winter EM, Abt GA, Nevill AM. Metrics of meaningfulness as opposed to sleights of significance: Taylor & Francis; 2014.
24. Field A. *Discovering statistics using IBM SPSS statistics*, Sage; 2013.
25. Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989:255-268.

26. Leo JA, Sabapathy S, Simmonds MJ, et al. The Respiratory Compensation Point is Not a Valid Surrogate for Critical Power. *Med Sci Sports Exerc*. 2017.

27. McBride G. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. *NIWA Client Report: HAM2005-062.* 2005.

28. Poole D. Measurements of the anaerobic work capacity in a group of highly trained runners. *Med Sci Sports Exerc*. 1986; 18(6):703-705.

29. Housh DJ, Housh TJ, Bauge SM. A methodological consideration for the determination of critical power and anaerobic work capacity. *Res Q Exerc Sport.* 1990; 61(4):406-409.

30. Jones AM, Vanhatalo A, Burnley M, et al. Critical power: implications for determination of VO2max and exercise tolerance. *Med Sci Sports Exerc.* 2010; 42(10):1876-1890.

31. McLellan TM, Cheung KS. A comparative evaluation of the individual anaerobic threshold and the critical power. *Med Sci Sports Exerc.* 1992; 24(5):543-550.

32. Mattioni Maturana F, Keir DA, McLay KM, et al. Critical power testing or self-selected cycling: Which one is the best predictor of maximal metabolic steady-state? *J Sci Med Sport.* 2017; 20(8):795-799.

33. Jones AM, Vanhatalo A. The 'Critical Power'Concept: Applications to Sports Performance with a Focus on Intermittent High-Intensity Exercise. *Sports Med.* 2017:1-14.

Figure 1.

Absolute difference (watt) between each proposed method (*see method section*) and the selected criterion ($CP_{3\text{-}hyp(1,2,3,4,5)}$) are presented along with 90% confidence interval around each difference along with the effect size calculation. The two vertical lines, placed along the x-scale, highlight the Smallest Worthwhile Change (±8 W) and represent the boundaries for a substantially negative or positive difference (displayed as grey areas). Numbers shown are the chances (%) that each difference is negative, trivial and positive. Inferences are also reported in qualitative terms as *almost certainly not*, *very unlikely*, *unlikely* or *probably not*, *possibly not*, *possibly*, *likely* or *probably, very likely* and *almost certain*.

Figure 2.

Absolute difference (Joules) between each proposed method (see method section) and the selected criterion ($CP_{3\text{-}hyp(1,2,3,4,5)}$) are presented along with 90% confidence interval around each difference along with the effect size calculation. The two vertical lines, placed along the x-scale, highlight the Smallest Worthwhile Change (± 1500 J) and represent the boundaries for a substantially negative or positive difference (displayed as grey areas). Numbers shown are the chances (%) that each difference is negative, trivial and positive. Inferences are also reported in qualitative terms as almost certainly not, very unlikely, unlikely or probably not, possibly not, possibly, likely or probably, very likely and almost certain.

Table 1. Group mean percent PO_{peak}, duration, absolute PO, and mechanical work during the five time-to-exhaustion trials for estimation of critical power.

Trial	%PO_{peak}	Duration (min)	Absolute PO (W)	Mechanical Work (kJ)
1	110 ± 0	1.7 ± 0.4	413 ± 60	42.7 ± 13.4
2	95 ± 0	3.2 ± 0.8	354 ± 59	66.5 ± 16.6
3	80 ± 0	7.1 ± 1.8	303 ± 50	127.9 ± 31.2
4	75 ± 4	12.5 ± 1.9	281 ± 47	209.8 ± 44.0
5	72 ± 4	19.4 ± 3.4	271 ± 46	312.1 ± 57.5

Table 2. Critical Power (CP) parameter estimates for the models CP_{exp}, $CP_{3\text{-}hyp}$, $CP_{2\text{-}hyp}$, CP_{linear}, and $CP_{1/time}$ using different combinations of time-to-exhaustion (TTE) trials.

		5	**4**		**3**				**2**			
			1,2,3,4	**2,3,4,5**	**1,2,3**	**1,3,5**	**2,3,4**	**3,4,5**	**1,2**	**1,5**	**3,4**	**4,5**
CP$_{exp}$												
	CP	275 ± 47	281 ± 47	270 ± 46								
	SEE	5.7 ± 2.1	8.0 ± 2.8	2.9 ± 1.9								
CP$_{3\text{-}hyp}$												
	CP	252 ± 44	250 ± 41	250 ± 43								
	SEE	3.4 ± 1.8	5.9 ± 5.8	4.5 ± 4.6								
	W'	23.1 ± 7.6	25.8 ± 12.6	24.9 ± 10.0								
	SEE	3.7 ± 2.0	5.9 ± 6.3	6.3 ± 7.3								
CP$_{2\text{-}hyp}$												
	CP	253 ± 44	256 ± 42	253 ± 44	263 ± 45	254 ± 43	256 ± 42	252 ± 44				
	SEE	1.6 ± 1.1	2.8 ± 2.2	2.3 ± 1.7	5.0 ± 6.9	1.6 ± 1.5	3.4 ± 3.2	2.6 ± 2.6				
	W'	20.3 ± 5.9	18.7 ± 6.7	20.1 ± 6.0	16.4 ± 5.7	19.8 ± 7.0	19.2 ± 7.1	21.2 ± 6.5				
	SEE	1.7 ± 1.0	1.9 ± 1.4	2.2 ± 1.5	1.6 ± 2.0	1.8 ± 1.5	2.3 ± 2.1	2.6 ± 2.3				
CP$_{linear}$												
	CP	256 ± 45	259 ± 44	255 ± 45	265 ± 47	256 ± 45	256 ± 45	253 ± 44	272 ± 50	257 ± 45	252 ± 46	251 ± 44
	SEE	2.3 ± 1.1	3.9 ± 2.7	2.4 ± 1.4	6.1 ± 6.5	3.1 ± 2.5	4.2 ± 3.5	2.4 ± 2.1				
	W'	17.9 ± 5.7	17.1 ± 5.7	19.2 ± 5.8	15.9 ± 5.7	17.7 ± 5.9	18.7 ± 5.9	21.0 ± 7.1	14.8 ± 6.5	16.3 ± 5.9	21.7 ± 8.8	22.3 ± 7.6
	SEE	1.5 ± 0.8	1.7 ± 1.1	1.8 ± 1.0	1.6 ± 1.5	2.4 ± 2.0	2.1 ± 1.8	2.0 ± 1.7				
CP$_{1/time}$												
	CP	261 ± 45	263 ± 45	256 ± 45	268 ± 47	260 ± 45	257 ± 45	253 ± 45	272 ± 50	257 ± 45	252 ± 46	251 ± 44
	SEE	4.4 ± 2.4	5.8 ± 3.4	3.1 ± 1.9	7.9 ± 5.6	4.6 ± 3.5	4.4 ± 3.6	2.4 ± 1.8				
	W'	16.1 ± 6.0	15.8 ± 5.9	18.4 ± 5.6	15.2 ± 6.0	16.0 ± 5.9	18.2 ± 5.7	21.1 ± 7.6	14.8 ± 6.5	16.3 ± 5.9	21.7 ± 8.8	22.4 ± 7.5
	SEE	0.8 ± 0.4	1.0 ± 0.6	1.0 ± 0.6	1.1 ± 0.8	0.8 ± 0.6	1.3 ± 1.0	1.5 ± 1.2				

CP_{exp}, exponential model; $CP_{3\text{-}hyp}$, 3-parameter hyperbolic model; $CP_{2\text{-}hyp}$, 2-parameter hyperbolic model; CP_{linear}, linear model; $CP_{1/time}$, 1/time linear model; CP, critical power; SEE, standard error of the estimation; W', anaerobic work capacity. Numbers on the top first and second row identify the number of trials and their corresponding durations (Table 1), respectively.

Figure 1
Click here to download high resolution image

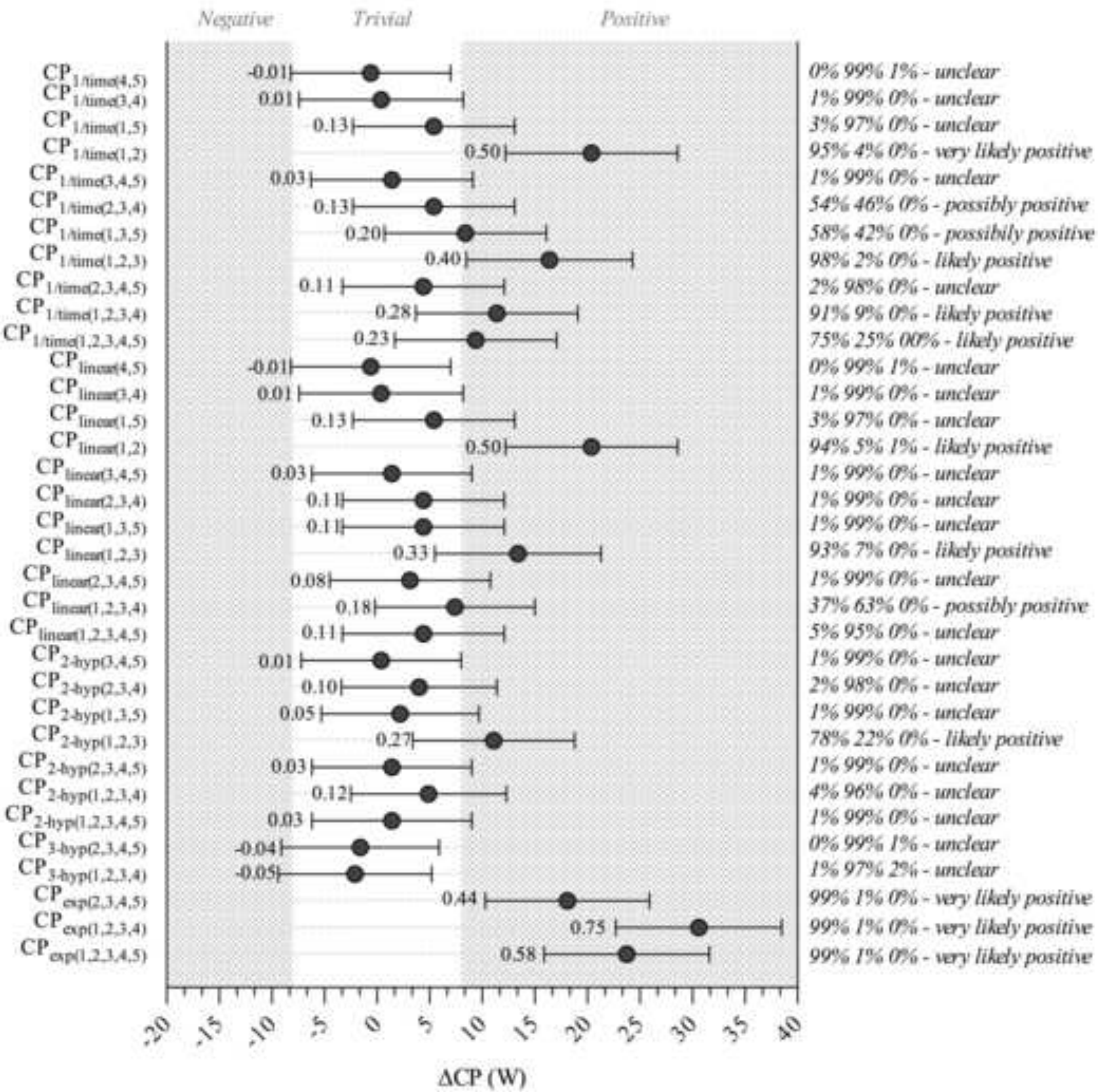
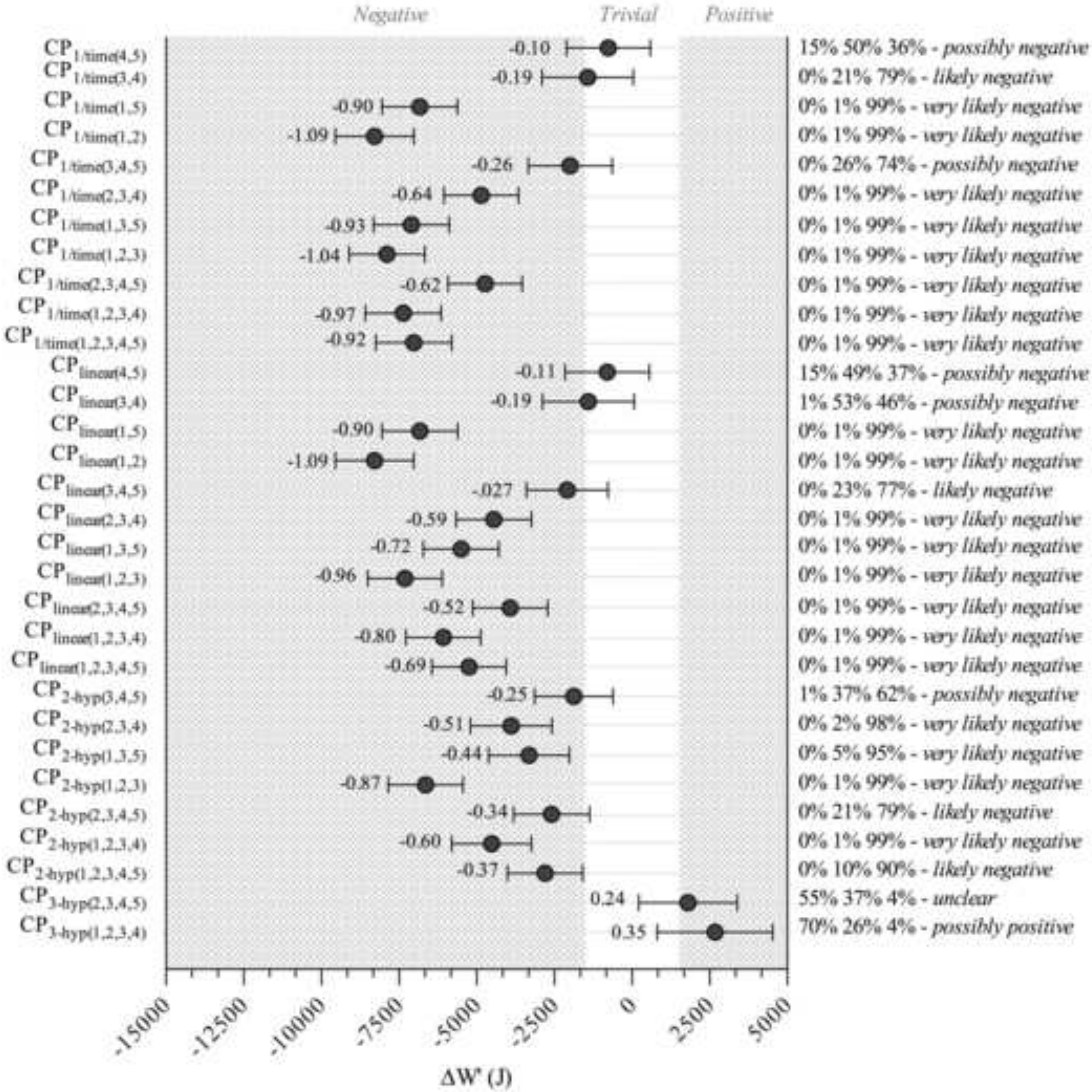


Figure 2

Click here to download high resolution image



Supplementary Material 1. Concordance correlation coefficient (CCC) analysis.

	#	R^2	RMSE (W)	CCC 95% CI [LL UL]	Slope (b_1) 95% CI [LL UL]	Intercept (b_0) 95% CI [LL UL]
$CP_{exp\ (1,2,3,4,5)}$	5	0.98	6.80	0.86[d] [0.75 0.97]	0.93 [0.84 1.02]	-5.47 [-31.10 20.10]
$CP_{exp\ (1,2,3,4)}$	4	0.96	9.70	0.79[d] [0.64 0.92]	0.92 [0.79 1.05]	-7.13 [-44.40 30.10]
$CP_{exp\ (2,3,4,5)}$	4	0.99	4.60	0.91[c] [0.84 0.98]	0.96 [0.89 1.02]	-6.73 [-23.90 10.50]
$CP_{3\text{-}hyp\ (1,2,3,4)}$	4	0.96	9.60	0.97[b] [0.94 1.00]	1.06 [0.91 1.22]	-14.39 [-52.50 23.70]
$CP_{3\text{-}hyp\ (2,3,4,5)}$	4	0.98	6.29	0.99[a] [0.98 1.00]	1.02 [0.93 1.11]	-3.51 [-27.10 20.20]
$CP_{2\text{-}hyp\ (1,2,3,4,5)}$	5	0.99	2.86	0.99[a] [0.99 1.00]	0.99 [0.95 1.03]	-0.10 [-10.60 10.40]
$CP_{2\text{-}hyp\ (1,2,3,4)}$	4	0.98	5.47	0.98[b] [0.97 1.00]	1.04 [0.96 1.12]	-15.91 [-37.30 5.48]
$CP_{2\text{-}hyp\ (2,3,4,5)}$	4	0.99	3.03	0.99[a] [0.99 1.00]	1.01 [0.96 1.04]	-2.03 [-13.24 9.16]
$CP_{2\text{-}hyp\ (1,2,3)}$	3	0.91	14.10	0.92[c] [0.84 1.00]	0.94 [0.74 1.14]	3.59 [-49.50 56.70]
$CP_{2\text{-}hyp\ (1,3,5)}$	3	0.99	3.01	0.99[a] [0.99 1.00]	1.01 [0.97 1.06]	-6.67 [-18.20 4.70]
$CP_{2\text{-}hyp\ (2,3,4)}$	3	0.98	5.47	0.98[b] [0.97 1.00]	1.05 [0.96 1.13]	-17.20 [-38.60 4.32]
$CP_{2\text{-}hyp\ (3,4,5)}$	3	0.99	3.51	0.99[a] [0.99 1.00]	0.99 [0.94 1.04]	0.72 [-12.10 13.60]
$CP_{linear\ (1,2,3,4,5)}$	5	0.99	3.60	0.99[a] [0.98 1.00]	0.98 [0.93 1.03]	-0.43 [-13.70 12.80]
$CP_{linear\ (1,2,3,4)}$	4	0.98	5.66	0.98[b] [0.95 1.00]	0.99 [0.91 1.10]	-4.45 [-25.60 16.75]
$CP_{linear\ (2,3,4,5)}$	4	0.99	2.44	0.99[a] [0.99 1.00]	0.98 [0.95 1.01]	0.34 [-8.60 9.30]
$CP_{linear\ (1,2,3)}$	3	0.93	11.98	0.92[c] [0.84 1.00]	0.90 [0.74 1.06]	12.75 [-30.20 55.70]
$CP_{linear\ (1,3,5)}$	3	0.99	3.57	0.99[a] [0.98 1.00]	0.98 [0.93 1.03]	-0.90 [-14.00 12.00]
$CP_{linear\ (2,3,4)}$	3	0.98	5.72	0.99[a] [0.97 1.00]	0.98 [0.90 1.06]	-0.18 [-21.30 20.90]
$CP_{linear\ (3,4,5)}$	3	0.99	2.44	0.99[a] [0.99 1.00]	0.99 [0.96 1.03]	0.53 [-8.40 9.40]
$CP_{linear\ (1,2)}$	2	0.75	22.88	0.78[d] [0.58 0.98]	0.77 [0.47 1.06]	42.55 [-37.60 122.60]
$CP_{linear\ (1,5)}$	2	0.99	3.97	0.99[a] [0.98 1.00]	0.98 [0.93 1.04]	-1.91 [-16.50 12.80]
$CP_{linear\ (3,4)}$	2	0.98	6.56	0.99[a] [0.98 1.00]	0.95 [0.86 1.04]	11.00 [-12.20 34.20]
$CP_{linear\ (4,5)}$	2	0.97	7.80	0.98[b] [0.97 1.00]	0.99 [0.88 1.10]	2.10 [-26.60 30.70]
$CP_{1/time\ (1,2,3,4,5)}$	5	0.97	7.50	0.96[b] [0.93 1.00]	0.97 [0.87 1.10]	-3.20 [-31.20 12.70]
$CP_{1/time\ (1,2,3,4)}$	4	0.96	9.13	0.94[c] [0.88 1.00]	0.96 [0.83 1.07]	-1.23 [-35.50 32.90]
$CP_{1/time\ (2,3,4,5)}$	4	0.98	5.53	0.98[b] [0.97 1.00]	0.97 [0.89 1.05]	2.28 [-17.50 22.50]
$CP_{1/time\ (1,2,3)}$	3	0.92	13.26	0.89[d] [0.79 0.99]	0.91 [0.72 1.08]	8.80 [-40.10 57.60]
$CP_{1/time\ (1,3,5)}$	3	0.98	7.10	0.97[b] [0.94 1.00]	0.97 [0.87 1.07]	-0.10 [-26.50 26.40]
$CP_{1/time\ (2,3,4)}$	3	0.97	8.43	0.98[b] [0.95 1.00]	0.96 [0.85 1.06]	3.80 [-27.10 34.70]
$CP_{1/time\ (3,4,5)}$	3	0.99	2.70	0.99[a] [0.99 1.00]	0.99 [0.96 1.03]	0.06 [-9.80 9.90]
$CP_{1/time\ (1,2)}$	2	0.75	22.90	0.78[d] [0.58 0.98]	0.76 [0.48 1.05]	43.20 [-36.70 123.23]
$CP_{1/time\ (1,5)}$	2	0.99	3.97	0.99[a] [0.98 1.00]	0.99 [0.93 1.04]	-1.90 [-16.60 12.80]
$CP_{1/time\ (3,4)}$	2	0.98	6.56	0.99[a] [0.98 1.00]	0.95 [0.86 1.04]	11.10 [-12.20 34.20]
$CP_{1/time\ (4,5)}$	2	0.97	7.80	0.98[b] [0.97 1.00]	0.99 [0.88 1.10]	2.10 [-26.60 30.70]

Relationships between the gold standard method ($CP_{3\text{-}hyp\ (1,2,3,4,5)}$) and all the other methods are expressed as R^2, root mean square error (RMSE), concordance correlation coefficient (CCC), slope and intercept. 95% CI around statistics are expressed as [LL UL]. [a], [b], [c], [d] represent the CCC descriptors namely almost perfect agreement (CCC > 0.99), substantial agreement (0.95 > CCC < 0.99), moderate agreement (0.90 < CCC > 0.95), and poor agreement (CCC < 0.90) respectively.

Acknowledgements

No sources of external funding were received for this work.

The authors declare no conflicts of interest.