Reliability of the parameters of the power-duration relationship using maximal effort time-trials under laboratory conditions

Christoph Triska[1¶*], Bettina Karsten[2,3¶], Bernd Heidegger[1¶], Bernhard Koller-Zeisler[1,5¶], Bernhard Prinz[4¶], Alfred Nimmerichter[4¶], Harald Tschan[1¶]

[1]Centre for Sport Science and University Sports, University of Vienna, Austria.

[2]Department of Life and Sport Science, University of Greenwich, Kent, United Kingdom.

[3]Department of Exercise and Sport Science, LUNEX International University of Health, Exercise and Sports, Differdingen, Luxemburg.

[4]Training and Sports Sciences, University of Applied Sciences, Wr. Neustadt, Austria.

[5]Austrian Institute of Sports Medicine, Vienna, Austria.

*Corresponding author

E-mail: christoph.triska@univie.ac.at (CT)

[¶]These authors contributed equally to this work

Abstract

The purpose of this study was to assess the reliability of critical power (CP) and the total amount of work accomplished above CP (*W´*) across repeated tests using ecological valid maximal effort time-trials (TTs) under laboratory conditions. After an initial incremental exercise test, ten well-trained male triathletes (age: 28.5 ± 4.7 yrs; body mass: 73.3 ± 7.9 kg; height: 1.80 ± 0.07 m; maximal aerobic power (MAP): 328.6 ± 41.2 W) performed three testing sessions (Familiarization, Test I and Test II) each comprising three TTs (12 min, 7 min and 3 min with a passive recovery of 60 min between trials). CP and *W´* were determined using a linear regression of power vs. the inverse of time (1/t) ($P = W´ \cdot 1/t + CP$). A repeated measure ANOVA was used to detect differences in CP and *W´* and reliability was assessed using the intra-class correlation coefficient (ICC) and the coefficient of variation (CoV). CP and *W´* values were not significantly different between repeated tests ($P = 0.171$ and $P = 0.078$ for CP and *W´*, respectively). The ICC between Familiarization and Test I was $r = 0.86$ (CP) and $r = 0.58$ (*W´*) and between Tests I and II it was $r = 0.94$ (CP) and $r = 0.95$ (*W´*). The CoV notably decreased from 4.1% to 2.6% and from 25.3% to 8.2% for CP and *W´* respectively. Despite the non-significant differences for both parameter estimates between the repeated tests, ICC and CoV values improved notably after the Familiarization trial. Our novel findings indicate that for both, CP and *W´* post familiarization ICC and CoV values indicated high reliability. It is therefore advisable to familiarize well-trained athletes when determining the power-duration relationship using TTs under laboratory conditions.

Keywords

Endurance, fatigue, performance testing, exercise tolerance

Introduction

A reliable determination of critical power (CP) and the total amount of work accomplished above CP until task failure (*W´*) has long been a question of interest. Whilst CP represents a work rate that can be sustained for a long time without a continuous loss of metabolic (e.g. pH, phosphocreatine) and systemic (blood lactate concentration, $\dot{V}O_2$) homeostasis [1], *W´* is an equivalent for a finite amount of work that can be accomplished above CP [2, 3]. The original determination of CP and *W´* requires 3 to 5 constant-power time-to-exhaustion trials (TTE) on a cycle ergometer at various power outputs (PO) that lead to exhaustion between 2 and 15 min [e.g. 4, 5-7]. However, these TTE have no known endpoint and therefore are not comparable to the task athletes are confronted with during competitions. Using TTE for the determination of CP and *W´* provides reliable results for CP (r = 0.90-0.96) [8-10]. However, *W´* consistently demonstrates a poorer reliability across repeated tests (r = 0.64-0.84) [8-10]. Importantly, it has been shown that small errors in exhaustive duration might bias the parameter estimates (in particular *W´*) [11, 12]. Therefore, TTE efforts should only be used with caution when trying to detect small training induced changes in an athlete's performance [13].

Fixed duration time-trials (TT) with a known endpoint are typically used when CP and *W´* are determined under field conditions [4, 6, 7, 12, 14]. These TT carry a higher ecological validity compared to TTE and TT are often seen as optimal to approximate "real-world" conditions [4-7, 14, 15]. TT are further suggested to have a high test-retest reliability [16, 17] also when compared to TTE efforts [4, 18]. Moreover trained athletes are commonly accustomed to TT type efforts as well as they are familiar with competitive situations. Therefore, TT should be preferred over TTE when constructing the power-duration relationship. Reflecting ecological validity, Hampson et al. [19] argued that athletes during TT efforts are able to change the intensity according to their perception of fatigue and external motivational cues. Whilst

potentially adding some variability to the measurement, as intensity fluctuates throughout [13], Jeukendrup and Currell [20] debated that pacing is an inherent strategic component of real world performance and for that reason this should be included in performance tests. The only recent work suggesting a superiority of performance using TTE was performed by PASSFIELD AND COAKLEY (2017). Comparing time-matched TTE with TT, a higher average PO for the 80% TTE resulted in significantly higher values for CP and significantly lower *W´* values to those established from the TT.

When using TT for the determination of CP and *W´*, Galbraith et al. [15] and Karsten et al. [7] demonstrated a high reliability for critical speed (the mode equivalent of CP in running) and CP respectively using ecologically valid TT efforts in the field (coefficient of variation [CoV] = 1.3-2.0% [15]; CoV = 2.2-2.5% [7]). However, similar to TTE efforts both studies also demonstrated a poor reliability for TT determined values of *W´* [7, 15] (CoV = 9.8-18.4% [15]; CoV = 46.0-46.7% [7]). Karsten et al. [7] speculated that differences in environmental conditions (e.g. terrain, cadence) as well as differences in the seating position resulted in this low level of reliability of *W´*, whilst Galbraith et al. [15] found that a familiarization session had increased the reliability of *W´*. In contrast, Black et al. [21] recently reported a close agreement for *W´* but not for CP values when work-matching TTE and TT efforts in the laboratory. However, the ecological validity of these TT can be questioned. Whilst providing a time endpoint, the TT used a fixed resistance, allowing participants only to modulate efforts via cadence [21]. Consistent with Black et al. [21], Triska et al. [12] found non-significant differences and a significant correlation in *W´* between TTE and TT running using time-matching TTE and TT efforts. However, a high intra-individual variation did not allow the interchangeable use of *W´*. As some researchers suggested *W´* to show a high day-to-day variation and as questions have been raised whether *W´* can be accurately determined using the power-duration relationship, *W´* continues to be a debate [22, 23].

When testing for CP and *W´*, even well-trained cyclists appear to require two familiarization sessions when using fixed-duration TT in the laboratory. This was demonstrated by Parker Simpson and Kordi [24] who demonstrated significantly lower CP values during testing sessions 1 and 2 compared to subsequent sessions. Interestingly, no differences were found for *W´* across all trials. This is supported by other investigations, which showed a smaller CoV after a familiarization session, indicating the importance of familiarization [14, 15]. Galbraith et al. [15] argued that altered pacing strategies can result in smaller CoV values post familiarization. The same authors [14] also stated that the CoV had further diminished in a follow-up study which suggests that participants, as a result of the earlier investigation, were experienced with the testing protocol. Importantly, these authors demonstrated a non-reliable *W´* (ICC $r = 0.75$ and CoV = 32.7%) even though participants were familiarized [14]. However, the duration of the respective predictive runs were not matched in the latter study, what has been shown to affect the parameter estimates [12]. It is therefore still unclear what caused the differences in *W´*.

The present study follows a recent investigation evaluating the validity of laboratory based TTs to determine CP and *W´* [5]. This study conducted both trial modes (i.e. TTE and TT) and accounted for the differences in environmental conditions by performing all efforts under controlled laboratory conditions. Results once more demonstrated a poor agreement for *W´*, whilst CP values demonstrated close to identical values. Karsten et al. [5] consequently speculated that modelling issues might affect the determination of *W´* when using TTE efforts (e.g. notably lower SEE compared to TT) and the researchers postulated the need of low SEEs for a high quality model [5].

To-date the reliability of TT determined CP values has not been demonstrated in the laboratory. Given present findings for *W´* [7, 12, 14], familiarization, controlled conditions, and matched durations of respective trials might provide some further insight into this apparent conundrum. The aim of this study therefore was to assess the reliability and potential learning effects when using highly ecologically valid TT efforts to determine CP and *W´* under controlled laboratory conditions. After familiarization we hypothesized non-significant differences for CP and *W´*, a small CoV, and high ICC values.

Material and Methods

Participants

Ten well-trained male triathletes (age: 28.5 ± 4.7 yrs; body mass: 73.3 ± 7.9 kg; height: 1.80 ± 0.1 m; maximal aerobic power (MAP): 329 ± 41 W) volunteered to participate in this study. All participants were involved in regular training and competition for at least three years on a national competition level and they were experienced in performing TT. Before entering the study, participants had to complete a health questionnaire and provided written informed consent after the nature and risks of the study had been explained. The ethics committee of the University of Vienna (#00216) approved all experimental procedures and the study was conducted in accordance with the *Declaration of Helsinki*.

Study design

The study followed a repeated laboratory test design where participants reported to the laboratory on four occasions separated by at least 72 h. A preliminary graded exercise test (GXT) was followed by three visits consisting of three TTs each. These TTs were between 3 and 12 min in duration and interspersed by 60 min passive rest to allow blood lactate [La] return

to baseline values and to alleviate the effects of altered $\dot{V}O_2$ uptake kinetics [5, 23]. Tests were performed at the same time of the day (± 2 h) in an air-condition controlled laboratory. Temperature and relative humidity were between 22-23°C and 45-55%, respectively. Participants were instructed to arrive at the laboratory in a fully hydrated state and to avoid strenuous exercise the day before. Participants were also required to refrain from food, caffeine and alcohol intake the preceding 3 h. For all tests a Cyclus2 ergometer (RBM Elektronics, Leipzig, Germany) was used where participants used their personal racing or TT bikes which was mounted to the ergometer. Participants were instructed to use the same personal bike for all visits. During all tests, participants were strongly verbally encouraged. Testing was completed within 3 weeks to avoid effects of training and detraining. During the time of testing participants trained for approximately 3 to 5 h per week in their off-season. The majority completed the tests within 12-13 days, with the exception of a single participant who completed the study within 16 days. However, in this single participant the GXT and the familiarisation session were done 7 days apart from the two CP-tests which were interspersed by 72 h.

Graded exercise test

A GXT was performed to determine MAP. After an unloaded cycling phase for 3 min, resistance was set to 100 W and was increased by 20 W every 3 min until volitional exhaustion. If the last work stage could not be fully completed MAP was calculated using the following equation of Kuipers et al. [25]:

$$MAP = Plast + (\frac{t}{180} \cdot 20) \quad (1)$$

where MAP is the maximum aerobic power (W), P_{last} is the last fully completed work stage (W) and t is the duration of the incomplete work stage (s).

TT to determine the power-duration relationship

Participants performed three identical tests to determine the power-duration relationship. The first test was used as a familiarization session and it was included in the analysis. The first test is consequently termed *Familiarization*, and the second and third test *Test I* and *Test II*, respectively. During the TTs participants were advised to produce the highest mean power output for 12, 7 and 3 min in that order [26] and to end the trial fully exhausted ('maximal TT effort') [5]. To replicate real-world TT participants cycled at their own preferred cadence and they were free to change cadence during the trials. Transitions from rest to work were with an increase of pedal cadence to the participants' own preferred value after a 3-min unloaded cycling phase. During the TT PO increased as a function of cadence and pedal force. To simulate real-world TT, participants used a self-selected pacing strategy and they were able to adjust gearing throughout by using the virtual gear changer mounted to the handlebars.

Estimation of CP and W'

Mean PO for each TT was plotted against the inverse-of-time using a linear regression where P is the mean power output (W), W' is the total amount of work accomplished above CP until task failure (J) and CP is the critical power (W):

$$P = W' \cdot \frac{1}{t} + CP \quad (2)$$

Least square modelling procedures were used to fit the parameter estimates. The y-intercept represents CP and the slope represents W'. The individual SEE was calculated for each participant and each parameter estimate in absolute and relative values. Nimmerichter et al. [27] demonstrated that the model power vs. the inverse of time provides notably lower SEE compared to other two parameter models and therefore this model was used.

Statistical analyses

After testing for normality using Shapiro-Wilk procedures, a repeated measure analysis of variance (ANOVA) was conducted to assess differences between the tests. If the assumption of sphericity had been violated ($P < 0.001$) the Greenhouse-Geisser correction have been used [28]. Significant main effects were followed-up by Bonferroni post-hoc procedures. Partial eta-squared (η_p^2) was used to provide an estimate of effect size of the ANOVA (small $\eta_p^2 = 0.01$; moderate $\eta_p^2 = 0.10$; large $\eta_p^2 = 0.25$). Effect size for the post-hoc tests was calculated using Cohen's *d* (small $d = 0.2$; moderate $d = 0.5$; large $d = 0.8$) [29].The intra-class correlation coefficient (ICC) and the coefficient of variation (CoV) were calculated using a spreadsheet [30]. An ICC >0.9 indicates *high* reliability, values >0.8 indicate *moderate* reliability, values >0.6 indicate *questionable* reliability, and values <0.6 indicate *poor* reliability of repeated tests. The coefficient of variation (CoV) was used to rate intra-individual variation. An upper limit of 5% [30] or 10% [31] is proposed to provide reliable results when repeating two tests. The Bland-Altman's method of 95% limits of agreement (LoA) assessed the agreement between repeated tests for CP and *W´* [32]. Pearson product moment correlation assessed the strength of the relationship between repeated tests. Statistical significance was accepted at $P < 0.05$. Before the beginning of the study an *a priori* power-analysis was conducted and revealed that totally 10 participants were required to detect a significant difference of 15 W and 3 kJ for CP and *W´*, respectively with a statistical power of >80% [33]. A difference of 15 W in CP and 3 kJ in *W´* would result in a calculated TT_{20min} time difference of <5% what is well within day-to-day variation [12].

Results

Table 1 represents results of *Familiarization*, *Tests I* and *II*, Table 2 illustrates data reporting reliability and agreement between repeated tests (Figs 1 and 2), and Table 3 reports the ICC

and CoV of individual TTs. Figs 1 and 2 illustrate the correlation of CP and W' between repeated tests. Between tests non-significant differences were found for CP ($F_{2,18}$ = 1.949; P = 0.171; η_p^2 = 0.178) and W' ($F_{2,18}$ = 2.951; P = 0.078; η_p^2 = 0.247). Significant differences were found for the absolute SEE for CP ($F_{2,18}$ = 10.847; P = 0.001; η_p^2 = 0.547) and W' ($F_{2,18}$ = 10.865; P = 0.001; η_p^2 = 0.547) and the relative SEE for CP ($F_{2,18}$ = 5.935; P = 0.001; η_p^2 = 0.549) and W' ($F_{2,18}$ = 5.428; P = 0.014; η_p^2 = 0.376). Bonferroni post-hoc procedures for the absolute SEE revealed significant differences between *Familiarization* and *Test I* for CP and W' (P = 0.042 and d = 1.20 for both parameters) and between *Familiarization* and *Test II* for CP and W' (P = 0.008 and d = 1.74 for both parameters). No significant differences were found for the absolute SEE for CP (P = 0.989 and d < 0.01) and the absolute SEE for W' (P = 0.945 and d < 0.01) between *Test I* and *Test II.* Bonferroni post-hoc procedures for the relative SEE revealed significant differences between *Familiarization* and *Test I* and between *Familiarization* and *Test II* for CP only (P = 0.043, d = 1.04 and P = 0.005, d = 1.85, respectively), but not for W'. No significant differences were found for the relative SEE for CP (P = 0.850 and d = 0.12) and the relative SEE for W' (P = 0.841 and d = 0.12) between *Test I* and *Test II.*

Table 1: Results of CP and W' and their associated SEE.

	Familiarization	*Test I*	*Test II*
CP (W)	294 ± 26	302 ± 28	304 ± 29
W' (J)	17316 ± 6340	14972 ± 3052	14710 ± 3368
SEE CP (W)	7.2 ± 3.4	3.8 ± 3.2*	3.1 ± 3.0*
SEE W' (J)	2012 ± 963	1060 ± 896*	868 ± 825*
SEE CP (%)	2.4 ± 1.1	1.3 ± 1.1*	1.0 ± 1.0*
SEE W' (%)	12.6 ± 7.4´	7.3 ± 6.5	6.0 ± 6.0

CP = Critical Power; W' = maximum work above CP; SEE = standard error of the estimate; *significantly different at P < 0.050 from *Familiarization*.

Table 2: ICC (95%CL), CoV (95%CL), mean bias and 95% LoA for W' and CP.

	W' (J)	CP (W)
ICC *Familiarization* vs. *Test I*	0.58 (-0.03 to 0.88)	0.86 (0.53 to 0.96)
ICC *Test I* vs. *Test II*	0.95 (0.80 to 0.99)	0.94 (0.78 to 0.98)

CoV (%) *Familiarization* vs. *Test I*	25.3 (16.8 to 50.9)	4.1 (2.8 to 7.7)
CoV (%) *Test I* vs. *Test II*	8.2 (5.6 to 15.5)	2.6 (1.8 to 4.8)
Bias *Familiarization – Test I*	2742	-8
95% LoA	-6899 to 12384	-40 to 24
Bias *Test I – Test II*	-135	-2
95% LoA	-2635 to 2366	-24 to 21

ICC = intra-class correlation coefficient; CL = confidence limits; CoV = coefficient of variation; LoA = limits of agreement.

Table 3: ICC (95% CL), and CoV (95% CL) for individual TT and test

	12-min TT	7-min TT	3-min TT
ICC Fam – Test I	0.94 (0.78-0.99)	0.97 (0.88-0.99)	0.95 (0.80-0.99)
CoV Fam – Test I	2.9 (2.0-5.3)	2.0 (1.3-2.6)	3.0 (2.0-5.5)
ICC Test I – Test II	0.95 (0.83-0.99)	0.95 (0.82-0.99)	0.97 (0.87-0.99)
CoV Test I – Test II	2.4 (1.6-4.4)	2.5 (1.7-4.6)	2.5 (1.7-4.6)

Fam = Familiarization

Fig 1. Relationships (panels a and b) and Bland-Altman plots of the differences (panels c and d) between repeated tests of CP. The black solid line represents the linear regression and the grey-dotted line represents the line of identity. The solid grey line represents the mean bias and the dotted black line represent the 95% limits of agreement. The right panels (b and d) clearly show improved reproducibility of the data after the familiarization.

Fig 2. Relationships (panels a and b) and Bland-Altman plots of the differences (panels c and d) between repeated tests of *W´*. The black solid line represents the linear regression and the grey-dotted line represents the line of identity. The solid grey line represents the mean bias and the dotted black line represent the 95% limits of agreement. The right panels (b and d) clearly show improved reproducibility of the data after the familiarization.

Discussion

The main findings of the present study were that both, CP and *W´* values provide reliable results in a cohort of well-trained athletes after a familiarization trial. Importantly, this is the first study, which demonstrates such a reliability for the estimates of *W´*. Even though participants were

familiar with TT efforts in the field, they produced slightly higher CP estimates (~3.5%) and notably lower *W´* estimates (~13%) after the familiarization trial. Although non-significant differences in the parameter estimates were revealed, the effect size is of a moderate order for both parameter estimates. Considering effect sizes to be more appropriate when assessing smaller sample sizes and small mean differences [34], *small* effects were observed between *Familiarization* and *Test I* for CP (d = 0.28) and *W´* (d = 0.47) evaluating post-hoc analysis. The effect sizes for CP and *W´* between *Tests I* and *II* were *trivial* (d = -0.04 and d = -0.06, respectively).

Results demonstrate a notable improvement for ICC and CoV values related to both parameter estimates after familiarization using TTs of equal duration (i.e. 12, 7, and 3 min). Recently, it was demonstrated [12] that the high intra-individual variation in parameter estimates can be reduced when using iso-duration TTs compared with TTE efforts. The predictive error of *W´* however, remained unacceptably high. Like Vandewalle et al [11], the researchers also suggested *W´* to be sensitive to small changes in TTE durations [12]. Consequently, using fixed-duration TTs can alleviate these negative influences thus increasing reliability of the parameter estimates.

ICCs for CP between *Familiarization* and *Test I* and between *Tests I* and *II* can be interpreted as *moderate* and *highly reliable*, respectively. The CoV for CP notably dropped after the familiarization trial (4.1% vs. 2.6%), but interestingly both testing trials were within what is currently accepted as an accepted range (i.e. <10% in [31] and <5% in [30]) throughout testing sessions. Our CP results are consistent with studies where reliability of CP was evaluated using TTs under laboratory conditions [24] and under field conditions [7]. Karsten et al. [7] found similar ICC and CoV compared to the present results (ICC r = 0.99 and CoV = 2.2%). A recent study by Wright et al. [28] found comparable ICC (r = 0.94-0.99 in [28]) and comparable CoV

(1.2% and 8.4% in for CP and *W´* respectively), when using the 3MT. However, whilst employing TTs for the determination of the parameter estimates is a valid method [5], the validity for the 3MT is poor [28], which suggests that the determination of the parameter estimates using multiple TTs provides more accurate parameter estimates compared to a single effort, i.e. the 3MT.

Interestingly, the ICC for *W´* is only *poor* between *Familiarization* and *Test I*, but changes to be *highly reliable* between *Tests I* and *II*. Furthermore, the CoV was >10% for *W´* between *Familiarization* and *Test I*, whilst it improved to values that according to Atkinson and Nevill [31] can be seen as reliable (i.e, <10%) between *Tests I* and *II*, confirming *W´* to be reliable post familiarization. However, such high reliability was not present in a field-based study using a similar methodology (ICC $r = 0.16$ and CoV = 46% in [7]). Karsten et al. [7] speculated that differences in environmental conditions (e.g. level vs. uphill) might have influenced the results for *W´*. With the exclusion of this factor, our laboratory-based parameter estimates demonstrate a high level of reliability after familiarization (ICC $r = 0.95$). It can therefore be suggested, that standardized and controlled laboratory conditions alleviate influencing effects on *W´* and consequently result in a higher reliability of the parameter estimate.

The mean bias of CP and *W´* between *Tests I* and *II* was close to zero after a familiarization session (Fig 1). Furthermore, the 95% LoA for both parameters showed notably closer LoA after *Familiarization* (Fig 1) which is consistent with findings using TTs in well-trained runners [15]. Galbraith et al. [15] improved their 95% LoA for *W´* from ±80 m to ±45 m (reduction of ~50%), and performing a familiarization session in the present study resulted in an even greater reduction of the 95% LoA (± 10,000 J to ± 2,500 J) (reduction of ~75%). These results suggest the evidence of a learning effect even in well-trained cyclists. Similar to the LoA, the SEE became notably smaller for both parameter estimates after a familiarization session (Fig 2). Our

participants were able to provide a more consistent performance thereby reducing SEE by ~30% (CP) and by ~50% (*W´*) after *Familiarization*, also showing the presence of a learning effect. After *Familiarization*, a high agreement of the regression line and the line of identity for both parameter estimates was evident (Fig 2b and Fig 2d). The SEEs between *Tests I* and *II* (±12 W and ±1.3 kJ for CP and *W´* respectively) are also within day-to-day variations and they are lower compared to the recent field-based study by Karsten et al. [7]. SEE for CP in our study is slightly higher compared to another laboratory-based investigation using TT, however, the SEE for *W´* is similar [24]. It is important to note that Parker-Simpson and Kordi [24] used a different testing methodology by performing the third TT on a different day.

Moreover, Black et al. [21] and Karsten et al. [6] speculated that changes in cadence, different pacing pattern (i.e. fast start vs. slow start) between efforts could affect the determination of CP and *W´*. Galbraith et al. [14] reported a pacing related learning effect in well-trained runners which might be the cause for the low reliability between *Familiarization* and *Test I* in the present study. However, participants seem to have adapted a reproducible pacing strategy after a single familiarization session as the mean PO within the first 60 s was not different between *Test I* and *Test II* ($P = 0.561$). Contrary to this, Parker-Simpson and Kordi [24] stated the need of two familiarization sessions using TTs, but in contrast to the present study, participants were not allowed to change gear ratios during the TTs, which lowered ecological validity of the efforts and likely added to a larger learning effect.

Individual mean TT PO across all trials showed a high reliability ($r = 0.94$-0.97) and a low CoV (2.0-3.0%) (Table 3). These ICC and CoV values are consistent with Laursen et al. [18] ($r = 0.88$-0.95 and CoV = 2.0-3.3%) who also argued for ecolgically valid TTs when evaluating performance as a high test reproducibility is needed to detect even small changes in an athlete's performance. Even though individual TTs were highly reliable throughout repeated tests,

notably lower SEE values (i.e. elevated quality of the model) after familiarization were demonstrated. Thereafter, SEE maintained low values in subsequent tests. The present results support the argument by Karsten et al. [5] who stated that assessing the SEE is an important measure for the quality of the model. The differences in absolute and relative SEE of CP and *W´* between *Familiarization* and *Test I* are of a large effect size, which shows a learning effect and consequently the need for familiarization. Recently, SEE values above recommended limits (i.e. 2% for CP and 10% for W´ [35, 36]) was suggested to bias the parameter estimates [5, 12]. Consequently, at least three TTs should be conducted to calculate SEEs.

Generally, the reasons for the higher reliability in the current study compared to earlier work could have been threefold: (i) controlled laboratory conditions; (ii) same TT durations across visits; (iii) no differences in pacing strategy after a familiarization session.

A potential limitation of the study was the use of fixed-duration TT. These whilst arguably carrying a higher ecological validity compared to constant-power TTE, are limited by competitive races commonly using fixed-distances rather than fixed-times. Still, fixed-time TT should be preferred as t_{lim} for each participant is equal reducing the level of random error and construct the power-duration relationship reproducibly [12]. More research can be suggested to investigate the potential supremacy of fixed-distance TT in the laboratory and the field.

Conclusion

To reduce the error inherent in testing, present results demonstrate that trained athletes experienced in TT and competitive events require to be familiarized when determining CP and *W´* using TT in the laboratory. Even though highly reliable results for individual mean TT PO across multiple tests were evident, the quality of the model increased in subsequent testing

sessions. Therefore, using TT is valid, reliable, and ecologically valid (i.e. own pacing strategy, change of cadence and gearing). We consequently suggest that laboratory TTs are preferable over TTE efforts and TT rather than TTE should be considered as a recommended method of best practice when determining CP and W'.

Acknowledgement

The corresponding author wants to express his sincere thanks to the Austrian Institute of Sports Medicine and the University of Applied Science Wr. Neustadt for providing equipment of their laboratories.

References

1. Jones AM, Wilkerson DP, DiMenna F, Fulford J, Poole DC. Muscle metabolic responses to exercise above and below the "critical power" assessed using 31P-MRS. Am J Physiol Regul Integr Comp Physiol. 2008;294(2):R585-93. doi: 10.1152/ajpregu.00731.2007. PubMed PMID: 18056980.

2. Hill DW. The critical power concept. A review. Sports Med. 1993;16(4):237-54. PubMed PMID: 8248682.

3. Moritani T, Nagata A, deVries HA, Muro M. Critical power as a measure of physical work capacity and anaerobic threshold. Ergonomics. 1981;24(5):339-50. doi: 10.1080/00140138108924856. PubMed PMID: 7262059.

4. Triska C, Tschan H, Tazreiter G, Nimmerichter A. Critical Power in Laboratory and Field Conditions Using Single-visit Maximal Effort Trials. Int J Sports Med. 2015;36(13):1063-8. doi: 10.1055/s-0035-1549958. PubMed PMID: 26258826.

5. Karsten B, Baker J, Naclerio F, Klose A, Bianco A, Nimmerichter A. Time Trials versus Time to Exhaustion Tests: Effects on Critical Power, W' and Oxygen Uptake Kinetics. Int J Sports Physiol Perform. 2017:1-22. doi: 10.1123/ijspp.2016-0761. PubMed PMID: 28530476.

6. Karsten B, Jobson SA, Hopker J, Jimenez A, Beedie C. High agreement between laboratory and field estimates of critical power in cycling. Int J Sports Med. 2014;35(4):298-303. doi: 10.1055/s-0033-1349844. PubMed PMID: 24022574.

7. Karsten B, Jobson SA, Hopker J, Stevens L, Beedie C. Validity and reliability of critical power field testing. Eur J Appl Physiol. 2015;115(1):197-204. doi: 10.1007/s00421-014-3001-z. PubMed PMID: 25260244.

8. Gaesser GA, Wilson LA. Effects of continuous and interval training on the parameters of the power-endurance time relationship for high-intensity exercise. Int J Sports Med. 1988;9(6):417-21. doi: 10.1055/s-2007-1025043. PubMed PMID: 3253231.

9. Nebelsick-Gullett LJ, Housh TJ, Johnson GO, Bauge SM. A comparison between methods of measuring anaerobic work capacity. Ergonomics. 1988;31(10):1413-9. doi: 10.1080/00140138808966785. PubMed PMID: 3208733.

10. Smith JC, Hill DW. Stability of parameter estimates derived from the power/time relationship. Can J Appl Physiol. 1993;18(1):43-7. PubMed PMID: 8471993.

11. Vandewalle H, Vautier JF, Kachouri M, Lechevalier JM, Monod H. Work-exhaustion time relationships and the critical power concept. A critical review. J Sports Med Phys Fitness. 1997;37(2):89-102. PubMed PMID: 9239986.

12. Triska C, Karsten B, Nimmerichter A, Tschan H. Iso-duration Determination of D' and CS under Laboratory and Field Conditions. Int J Sports Med. 2017;38(7):527-33. doi: 10.1055/s-0043-102943. PubMed PMID: 28514809.

13. Hinckson EA, Hopkins WG. Reliability of time to exhaustion analyzed with critical-power and log-log modeling. Med Sci Sports Exerc. 2005;37(4):696-701. PubMed PMID: 15809572.

14. Galbraith A, Hopker J, Lelliott S, Diddams L, Passfield L. A single-visit field test of critical speed. Int J Sports Physiol Perform. 2014;9(6):931-5. doi: 10.1123/ijspp.2013-0507. PubMed PMID: 24622815.

15. Galbraith A, Hopker JG, Jobson SA, Passfield L. A novel field test to determine critical speed. J Sport Medic Doping Studie. 2011;01(01):1-4.

16. Jeukendrup AE, Saris WH, Brouns F, Kester AD. A new validated endurance performance test. Med Sci Sports Exerc. 1996;28(2):266-70. PubMed PMID: 8775164.

17. Hopkins WG, Schabort EJ, Hawley JA. Reliability of power in physical performance tests. Sports Med. 2001;31(3):211-34. doi: Doi 10.2165/00007256-200131030-00005. PubMed PMID: 11286357.

18. Laursen PB, Francis GT, Abbiss CR, Newton MJ, Nosaka K. Reliability of time-to-exhaustion versus time-trial running tests in runners. Med Sci Sports Exerc. 2007;39(8):1374-9. doi: 10.1249/mss.0b013e31806010f5. PubMed PMID: 17762371.

19. Hampson DB, St Clair Gibson A, Lambert MI, Noakes TD. The influence of sensory cues on the perception of exertion during exercise and central regulation of exercise performance. Sports Med. 2001;31(13):935-52. PubMed PMID: 11708402.

20. Jeukendrup AE, Currell K. Should time trial performance be predicted from three serial time-to-exhaustion tests? Med Sci Sports Exerc. 2005;37(10):1820; author reply 1. PubMed PMID: 16260987.

21. Black MI, Jones AM, Bailey SJ, Vanhatalo A. Self-pacing increases critical power and improves performance during severe-intensity exercise. Appl Physiol Nutr Metab. 2015;40(7):662-70. doi: 10.1139/apnm-2014-0442. PubMed PMID: 26088158.

22. Galbraith A, Hopker J, Passfield L. Modeling Intermittent Running from a Single-visit Field Test. Int J Sports Med. 2015;36(5):365-70. doi: 10.1055/s-0034-1394465. PubMed PMID: 25665002.

23. Karsten B, Hopker J, Jobson SA, Baker J, Petrigna L, Klose A, et al. Comparison of inter-trial recovery times for the determination of critical power and W' in cycling. J Sports Sci. 2017;35(14):1420-5. doi: 10.1080/02640414.2016.1215500. PubMed PMID: 27531664.

24. Parker Simpson L, Kordi M. Comparison of Critical Power and W' Derived from Two or Three Maximal Tests. Int J Sports Physiol Perform. 2016:1-24. doi: 10.1123/ijspp.2016-0371. PubMed PMID: 27918663.

25. Kuipers H, Verstappen FT, Keizer HA, Geurten P, van Kranenburg G. Variability of aerobic performance in the laboratory and its physiologic correlates. Int J Sports Med. 1985;6(4):197-201. doi: 10.1055/s-2008-1025839. PubMed PMID: 4044103.

26. Jenkins DG, Quigley BM. Endurance training enhances critical power. Med Sci Sports Exerc. 1992;24(11):1283-9. PubMed PMID: 1435180.

27. Nimmerichter A, Steindl M, Williams CA. Reliability of the Single-Visit Field Test of Critical Speed in Trained and Untrained Adolescents. Sports. 2015;3(4):358-68. doi: 10.3390/sports3040358.

28. Wright J, Bruce-Low S, Jobson SA. The Reliability and Validity of the 3-min All-out Cycling Critical Power Test. Int J Sports Med. 2017;38(6):462-7. doi: 10.1055/s-0043-102944. PubMed PMID: 28388783.

29. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J.: L. Erlbaum Associates; 1988. xxi, 567 p. p.

30. Hopkins WG. A new view on statistics: Internet Society for Sport Science; 2000 [updated 17 August 2011]. Available from: http://www.sportsci.org/resource/stats/.

31. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 1998;26(4):217-38. PubMed PMID: 9820922.

32. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10. PubMed PMID: 2868172.

33. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods. 2009;41(4):1149-60. doi: 10.3758/BRM.41.4.1149. PubMed PMID: 19897823.

34. Buchheit M. The Numbers Will Love You Back in Return-I Promise. Int J Sports Physiol Perform. 2016;11(4):551-4. doi: 10.1123/IJSPP.2016-0214. PubMed PMID: 27164726.

35. Ferguson C, Wilson J, Birch KM, Kemi OJ. Application of the speed-duration relationship to normalize the intensity of high-intensity interval training. PLoS One. 2013;8(11):e76420. doi: 10.1371/journal.pone.0076420. PubMed PMID: 24244266; PubMed Central PMCID: PMCPMC3828304.

36. Dekerle J, de Souza KM, de Lucas RD, Guglielmo LG, Greco CC, Denadai BS. Exercise Tolerance Can Be Enhanced through a Change in Work Rate within the Severe Intensity Domain: Work above Critical Power Is Not Constant. PLoS One. 2015;10(9):e0138428. doi: 10.1371/journal.pone.0138428. PubMed PMID: 26407169; PubMed Central PMCID: PMCPMC4583487.