

# Reliability of a laboratory-based long sprint cycling test: applications of the smallest worthwhile changes in performance for repeated measures designs

## Confiabilidade de um teste em laboratório de sprint longo no ciclismo: aplicações da menor mudança significativa em desempenho para desenhos de medidas repetidas

Rogério Santos de Oliveira Cruz<sup>1</sup>  
Rafael Alves de Aguiar<sup>1</sup>  
Tiago Turnes<sup>1,2</sup>  
Felipe Domingos Lisboa<sup>1</sup>  
João Antônio Gesser Raimundo<sup>1</sup>  
Fabrizio Caputo<sup>1</sup>

**Abstract** – The aims of the present study were to assess the reliability of long sprint cycling performance in a group of recreationally trained cyclists and to provide thresholds for changes in performance for this particular group of subjects in repeated measures designs through a scale of magnitudes. Repeatability of mean power output during a 1-min cycling time trial was assessed in a group of 15 recreationally trained cyclists ( $26 \pm 5$ , years,  $176 \pm 5$  cm,  $78 \pm 8$  kg). They were tested on separate days, approximately one week apart. The test and retest values for the whole group of cyclists were  $7.0 \pm 0.5$  W/kg and  $6.9 \pm 0.6$  W/kg (systematic change and 90% confidence limits of  $-1.0\% \pm 1.1\%$ ). Our results indicated good test-retest reproducibility (typical error of 1.8%, 90% confidence limits of 1.4% to 2.6%; intraclass correlation coefficient of 0.96, confidence limits of 0.91 to 0.99), but suggested a reduction of mean power for the “slower” subjects on retest ( $-2.0\%$ , 90% confidence limits of  $\pm 1.8\%$ ). If not monitored, this systematic decrease could interfere in results of studies utilizing groups with similar performance levels, particularly investigating strategies to improve performance in sprint cycling exercises around 1 min. The thresholds for moderate, large, very large and extremely large effects for mean power output on long sprint cycling performance are about 0.4%, 1.3%, 2.3%, 3.6%, and 5.8%, respectively.

**Key words:** Physical education and training; Physical fitness; Reproducibility of results; Exercise test.

**Resumo** – Os objetivos do presente estudo foram avaliar a confiabilidade do desempenho de sprint longo no ciclismo em um grupo de ciclistas recreacionais e fornecer limiares para mudanças no desempenho para este grupo de sujeitos em delineamentos de medidas repetidas, através de uma escala de magnitudes. A reprodutibilidade da potência média durante um contrarrelógio de 1-min foi avaliada em um grupo de 15 ciclistas recreativamente treinados ( $26 \pm 5$  anos,  $176 \pm 5$  cm,  $78 \pm 8$  kg). Eles foram avaliados em dias separados, com aproximadamente uma semana de intervalo. Os valores de teste-reteste no contrarrelógio de 1-min para o grupo inteiro foram  $7,0 \pm 0,5$  W/kg and  $6,9 \pm 0,6$  W/kg (mudança sistemática e limite de confiança de 90% de  $-1,0\% \pm 1,2\%$ ). Nossos resultados indicam boa reprodutibilidade teste-reteste (erro típico de 1,8%, limite de confiança de 90% de 1,4% para 2,6%; coeficiente de correlação intraclasses de 0,96, limite de confiança de 90% de 0,91 para 0,99), mas sugeriu uma redução da média de potência para os sujeitos “mais lentos” no reteste ( $-2,0\%$ , limite de confiança de 90% de  $\pm 1,8\%$ ). Se não monitorada, esta queda sistemática na potência pode interferir nos resultados de estudos utilizando grupos com níveis similares de desempenho, particularmente investigando estratégias para melhorar o desempenho em sprints com duração próxima a 1 min. Os limiares para efeitos moderado, grande, muito grande e extremamente grande para média da potência em sprint longo no ciclismo são aproximadamente 0,4%, 1,3%, 2,3%, 3,6% e 5,8%, respectivamente.

**Palavras-chave:** Aptidão física; Educação física e treinamento; Reprodutibilidade dos testes; Teste de esforço.

1 Santa Catarina State University. College of Health and Sport Science. Human Performance Research Group. Florianópolis, SC. Brazil.

2 Federal University of Santa Catarina. Sports Center. Physical Effort Laboratory. Florianópolis, SC. Brazil.

Received: October 03, 2017  
Accepted: January 08, 2018



Licença  
Creative Commons

## INTRODUCTION

The performance of a subject during an intense exercise always shows random variation from trial to trial. For competitive athletes performing a long sprint cycling (e.g. 1-km sprint), enhancements or impairments of performance affect the chances of a medal only if they are greater than about half the magnitude of this random variation<sup>1-3</sup>. However, plenty of research in the field of sports performance is conducted with groups of non-athletes – employing usually from 48-h to one week recovery between tests – in an attempt to link changes in performance to changes in metabolic or physiological markers induced by exercise e.g.<sup>4,5</sup>. Although widely used in scientific research investigating the acute effects of a strategy in performance, for this particular group of subjects (i.e. recreationally trained cyclists), reliability studies on performance are scarce, notably for long sprint cycling.

As a poor reliability affects test sensitivity<sup>3</sup>, reproducibility has important implications for the sample size needed to infer over a given effect in straightforward crossover trials. In addition, the typical variation in performance between tests, which is a measure of reliability, is a benchmark for assessing the smallest important enhancements for performance. Accordingly, the concept of calculating the increase in the chances of winning an event after an intervention can be extrapolated to non-athletes participating in a repeated measures design. Hopkins et al.<sup>2</sup> showed through simulations that the increase in the chances of winning varies uniformly when a particular athlete is benefited with an enhancement corresponding to multiples of the within-subject random variation in a group of identical athletes. If this is so, this notion can be used to estimate the chances of subjects winning or losing when competing against themselves in hypothetical events, after being submitted to different treatments. Therefore, a reliability study allows for the development of a scale of magnitudes for changes in performance during long sprint cycling<sup>6</sup>.

The aims of the present study were 1) to assess the reliability of mean power output during long sprint cycling in a group of recreationally trained cyclists and 2) to provide thresholds for changes in performance through a scale of magnitudes. A 1-min sprint cycle test was used as the criterion exercise, the duration being similar to that of world-class men's 1-km track cycling<sup>7</sup>.

## METHODOLOGICAL PROCEDURES

### Ethical approval

The study was carried out in accordance with the guidelines contained in the Declaration of Helsinki and was approved by the Ethics Committee in Human Research of the Santa Catarina State University (246.876). The subjects were fully informed of any risks and discomforts associated with the experiments before giving their written informed consent to participate.

A group of 15 recreationally trained cyclists<sup>8</sup> (age,  $26 \pm 5$  years; height,  $176 \pm 5$  cm; body mass,  $78 \pm 8$  kg), volunteered for this study. None of whom were receiving any pharmacological or specific dietetic treatment. All participants attended properly fed and hydrated and were instructed not to perform strenuous exercise and to abstain from alcohol in the day before each session. They were also asked to maintain the same dietary pattern throughout the experiment and to refrain from consuming caffeine for at least 2-h before each trial.

## Study design

This study is part of a straightforward crossover trial looking at the effects of ischemic preconditioning of the lower limbs on long sprint cycling performance, which was published elsewhere<sup>9</sup>. Subjects were required to report to the laboratory on six occasions over a 15-day period ( $\pm 4$  days), and all tests were interspersed with  $\sim 48$  h of recovery. After an incremental test and a visit aiming familiarization with the 1-min time trial (sessions 1 and 2), subjects were randomly submitted in sessions 3 and 4 to a performance protocol preceded by either intermittent bilateral cuff inflation to 220 mm Hg or to 20 mm Hg (unable to modify the arterial inflow, i.e. control). To increase data reliability, the latter visits were replicated in visits 5 and 6, also in a random manner. For the purposes of the present study, the experimental conditions were discarded and only the control conditions were used. Each subject was always tested at the same time of day to minimize the effects of diurnal biological variation ( $\pm 2$  h) in a temperature-controlled laboratory ( $21 \pm 1^\circ\text{C}$ ). All cycle tests were performed on an electrodynamically braked cycle ergometer (Lode Excalibur Sport, Groningen, The Netherlands). The ergometer seat and handlebar were adjusted for comfort, and the settings were replicated for subsequent tests.

## Exercise protocol

In the first visit, subjects underwent an incremental test to determine maximal oxygen uptake, peak power output, and the intensity associated with the first lactate threshold. The test consisted of 3 min of unloaded baseline pedaling ( $16 \pm 4$  W, equivalent to the lowest workload provided by the equipment) followed by an increase in power output of 0.5 W per kg of body mass by every third minute. Subjects were instructed to maintain their preferred cadence for as long as possible until volitional exhaustion. The intensity associated with the lactate threshold was that immediately prior to the increase in lactate concentration above baseline. Peak power output was defined as the power output attained at exhaustion if the test was terminated at the end of a 3-min stage. If the test was terminated before the last stage had been completed, peak power was calculated as the power of the previous stage plus the power increment times the duration of exercise at the final stage (s) divided by 180 s. On the following session, subjects carried out a 1-min performance test for familiarization purposes. No measures were taken during this test.

Before the sprint tests, subjects were submitted to a moderate warm-up protocol consisting of two 6-min steps at 90% of the lactate threshold intensity. Both transitions were preceded by 3-min of unloaded baseline pedaling and interspersed with 5-min of passive recovery on the bike. Five minutes after the warm-up, subjects performed a 1-min seated sprint cycle test. The resistance applied on the pedals was that corresponding to 7.5% of the individual body weight. The participants commenced tests from stationary start after a 10-s countdown with the crank for their preferred leg positioned at 45° angle to the horizontal<sup>10</sup>. During the sprint, they were informed of the time elapsed every 10 s, but were unable to see the display of the ergometer and were not informed of their performance at any stage until the end of the experimental protocol. During all tests, subjects were always verbally encouraged to give their best effort and mechanical power output was continuously measured at a sampling rate of 5 Hz. Mean power was latter calculated.

### Statistical analysis

Calculations were performed with the aid of a spreadsheet for assessment of retest reliability<sup>11</sup>. Mean power values were log transformed for the analysis, because this approach yields variability as a percent of the mean (coefficient of variation; CV), which is the natural metric for most measures of athletic performance<sup>1, 12</sup>. Within- and total between-cyclist CV were derived by back transformation of the residuals returned by the spreadsheet. The pure between-subject CV (i.e. free of typical error) was derived by the square root of the difference between the total variance and the internal variance (i.e. the typical error). The three measures of reliability were the change in the mean, typical error of measurement, and intraclass correlation coefficient (ICC). The uncertainties in the effects were always expressed as 90% confidence limits. As potential heterogeneous responses were detected between the top and bottom cyclists, an analysis of two subgroups of cyclists based on their performance (relativized by body mass) was included (see Results). The typical variation in performance of the two subgroups of cyclists was compared by calculating the ratio of the within-cyclist CV, and deriving likely limits for the ratio via the t distribution<sup>13</sup>.

The calculations of the thresholds for performance effects, including the smallest worthwhile enhancement, were done by multiplying the typical error by the factors provided by Hopkins et al.<sup>6</sup>. The minimum sample sizes needed to detect intervention-induced effects of each size in straightforward crossover trials were also approximated. These calculations were made with the aid of a spreadsheet for sample-size estimation for magnitude-based inferences<sup>14</sup>. Acceptable Type I and II error rates were set as 0.5% and 25%, respectively.

## RESULTS

During the incremental test, subjects attained a peak power output of 299 ± 33 W, whereas the lactate threshold intensity was 89 ± 19 W.

The mean group power profile during the 1-min time trial was represented in Figure 1. The reproducibility statistics for the performance measurements are presented in Table 1. The random variation from test to test was apparently lower in the subgroup with better performance (ratio of CV of top/bottom halves, 0.8). However, for the few subjects within each subgroup, the confidence intervals of this ratio (0.4 to 1.5) were insufficient to conclude whether differences are real or simply result of sampling error. The high intraclass correlation coefficients means that the reproducibility of the rank order of subjects was maintained, i.e. in general cyclists maintained the same positions within the group/subgroup on retest.

**Table 1.** Between and within subject variability of all subjects and for two subgroups based on mean performance during the 1-min time trial.

		Top	Bottom	All
Number of cyclists		7	8	15
Mean Power $\pm$ SD (W/kg)	Test	7.5 $\pm$ 0.3	6.6 $\pm$ 0.3	7.0 $\pm$ 0.5
	Retest	7.5 $\pm$ 0.4	6.5 $\pm$ 0.4	6.9 $\pm$ 0.6
Systematic change and 90%CL (%)		0.0 $\pm$ 1.5	-2.0 $\pm$ 1.8	-1.0 $\pm$ 1.1
Total CV	(%)	4.7	5.1	9.0
	90%CL	(3.3 – 9.0)	(3.6 – 9.2)	(6.9 – 13.1)
Between-subject CV	(%)	4.5	4.7	8.8
	90%CL	(3.1 – 8.6)	(3.3 – 8.5)	(6.8 – 12.8)
Within-subject CV	(%)	1.4	1.9	1.8
	90%CL	(1.0 – 2.8)	(1.3 – 3.5)	(1.4 – 2.6)
Intraclass correlation	ICC	0.95	0.90	0.96
	90%CL	(0.78 – 0.99)	(0.67 – 0.97)	(0.91 – 0.99)

CV, coefficient of variation; ICC, intraclass correlation coefficient; 90%CL, likely range for the true value of the CV.

The thresholds for change in performance during the 1-min time trial are presented in Table 2, along with a scale for the magnitude of effects associated with the required sample size for crossover studies. Based on these estimates, there was probably a substantial decrease in the grand mean at retest (77% chances). However, although the limited number of subjects in each subgroup (and therefore a wide confidence interval) do not allow for a categorical statement on the lack of performance changes in the “faster” subgroup, the likely negative change between trials was highly affected by the subset of “slower” subjects (particularly by two specific subjects). They presented 91% changes of a meaningful decrease on the second trial. A careful analysis of the residuals in Figure 2 may assist in this understanding.

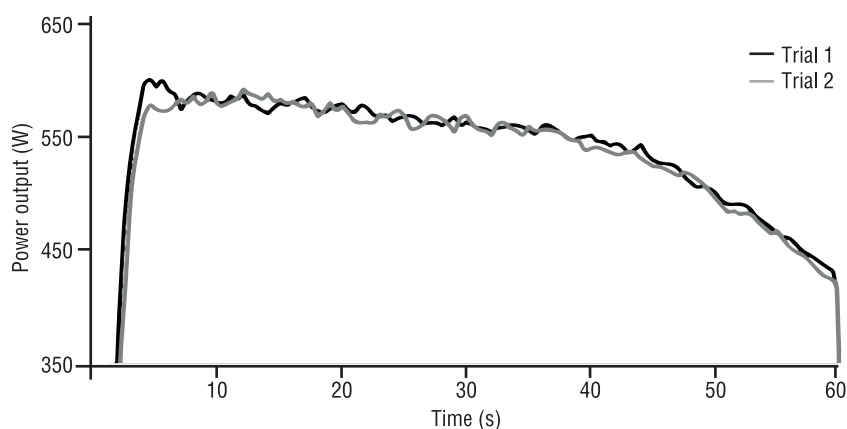
## DISCUSSION

The objectives of this study were 1) to analyze the reliability of a laboratory-based long sprint cycling test for a group of recreationally trained cyclists and 2) to identify how large the performance improvement resulting from a given intervention has to be to induce meaningful changes on athletic

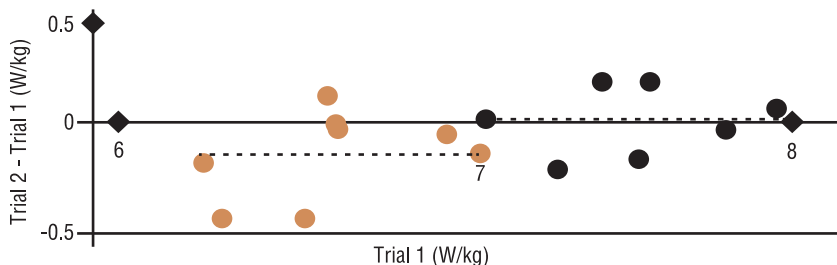
**Table 2.** A scale of magnitudes for changes in mean power during the 1-min time trial.

Group	Effect size (%)				
	Small <sup>a</sup>	Moderate	Large	Very large	Extremely large
Top 7	0.4 (0.3 – 0.8)	1.3 (0.9 – 2.5)	2.3 (1.6 – 4.4)	3.6 (2.5 – 7.0)	5.8 (4.0 – 11.1)
	62 (19 - 127)	10 (5 – 17)	6 (2 – 7)	2 (1 – 5)	1 (1 – 2)
Bottom 8	0.6 (0.4 – 1.0)	1.7 (1.2 – 3.1)	3.1 (2.1 – 5.5)	4.8 (3.4 – 8.6)	7.6 (5.4 – 13.8)
	37 (14 – 71)	8 (3 – 11)	3 (1 – 5)	1 (1 – 3)	1 (1 – 1)
All	0.5 (0.4 – 0.8)	1.6 (1.2 – 2.4)	2.9 (2.2 – 4.2)	4.5 (3.5 – 6.6)	7.2 (5.5 – 10.6)
	41 (21 – 67)	8 (5 – 10)	4 (2 – 5)	1 (1 – 2)	1 (1 – 1)

a Smallest worthwhile change for performance. Values in lower cases approximate the sample sizes needed to detect the associated effect sizes in straightforward crossover trials. Values in parentheses are 90%CL. Acceptable Type I and II error rates were set as 0.5% and 25%, respectively.



**Figure 1.** Group mean power profile during the 1-min time trial. Standard deviations have been omitted to preserve clarity.



**Figure 2.** Repeatability of mean power during the 1-min time trial (residual vs. predicted analysis). The dashed lines represent the mean performance change for each subgroup. Note the better consistency of the top 7 cyclists (closed symbols).

performance. Our results indicated good test-retest reproducibility for the whole group, but suggested a reduction of mean power for the slower subjects on retest. If not monitored, this systematic decrease could interfere in results of studies utilizing groups with similar performance levels, particularly investigating strategies to improve performance in sprint cycling exercises around 1-min.

A systematic change in the mean is a non-random change in the value between two trials, and can be caused by changes in the subjects' behavior during the experimental period<sup>12</sup>. In the present study, the mean absolute difference between test and retest was 5.6 W, which would result in roughly eight-tenths of a second if performance was based on a fixed distance (or



work) instead of a fixed time<sup>15</sup>. Although it has not been explicit during visits, some factors could have influenced these results; among them, the more likely is accumulated fatigue. Our performance protocol also involved a constant test until exhaustion at 100% of peak power output approximately one hour after the completion of the sprint. Therefore, it is possible that the recovery time between tests had been insufficient to allow complete recovery for a few subjects, leading to a slight reduction in the average power during the second trial. As only two subjects of the specific subgroup of slower cyclists appeared to be affected, i.e. those most susceptible to cumulative fatigue, other aspects as teleoanticipation and loss of intrinsic motivation after the first trial are less likely candidates. To avoid these undesirable effects in future experiments, it would be interesting to increase the time intervals between test sessions and/or to perform enough visits until any systematic change become negligible<sup>12</sup>.

The fitness level of our recreational cyclists (i.e. ability to generate power) proved to be quite homogeneous within the group. The pure between-subject CV was similar to those of specifically trained athletes competing in the 1-km time trial (a combination of professional and amateur riders)<sup>1</sup>. These unexpected similarities can be partially explained by our experimental design. As opposed of an actual competitive event, our subjects performed in a controlled laboratory environment, where all tests were held on the same cycle ergometer with the same crank arm length (and in some cases with the same cycling clipless shoes). On the other hand, while environmental conditions will have little effect in the 1-km time trial, because it is held in the more stable conditions of a velodrome<sup>1</sup>, features like air resistance, differences in equipment used by cyclists, and handling the bike through the course can constitute important sources of variation for athletes.

Indeed, compared with our subjects, trained cyclists appear to have somewhat lower values of external variation when the 1-min or 1-km time trials are performed in the laboratory<sup>7,15</sup>. It is important to note that the between-subjects CV should be taken into consideration to make a decision about subjecting or not an athlete to a given intervention, as the greater the dispersion in the cyclist's ability the greater must be the size of the improvement for the athlete to get important positions in the competition. However, this is not an important issue for non-athletes involved in a repeated measures design, once the determination of the smallest substantial improvement and the scale of magnitudes relies only on the typical error. In this sort of analysis, the chances of winning or losing after a given treatment are calculated by extrapolation to hypothetical competitions where the subject races against himself. Nevertheless, the external variation is pivotal for clustering subjects into a group of similar performance levels<sup>8</sup>, yielding similar metabolic response characteristics.

The analysis of two seasons combined (1999 and 2000), each comprised by three races over ~120 days, revealed that athletes of 1-km time trial (average performance of 69 s) present a within-subject CV of 1.0%

for performance times (90%CL of 0,8 – 1,4%)<sup>1</sup>. However, the equivalent within-subject CV to average power is approximately two to three times this value<sup>6</sup>, and therefore is possibly higher than that seen in the present study. Some characteristics may help explain the low internal variation of our cyclists in relation to athletes. In our work, the performance was repeated within a period of about a week, and all subjects were instructed to reproduce the same pacing strategy in all tests. Conversely, the performance data of the athletes were collected over a period of two years and possible sources of internal variation such as injuries or changes in pacing strategy could not be controlled. Moreover, as mentioned earlier, our subjects used the same ergometer in all tests, whereas the athletes can adapt/change their bicycles or even present mechanical problems within or between seasons. Thus, systematic changes in pacing strategy, in the skills of each individual, or in equipment throughout the seasons can help increase the typical error, being probably the factors influencing the higher variation observed in this population compared to our recreational cyclists.

One goal of this study was to identify how big the performance improvement resulting from a given intervention has to be to induce meaningful changes on the 1-min mean power of a recreationally trained cyclist. Hopkins et al.<sup>2</sup> showed that the increase in chances of victory varies uniformly when a particular subject is benefited with an effect corresponding to multiples of the intra-subject random variation in a group of identical individuals (i.e. between-subjects CV equal to zero, similar to a repeated measures design). Two assumptions are implicit in these simulations: subjects compete as independent individuals, and subjects attempt to perform their best in each trial<sup>1</sup>. By the possibility of overestimation of the group's typical error by the probable cumulative fatigue of some cyclists of the "slower" subgroup, we will focus on the intra-cyclist variation of the subgroup of better performance, as we believe that this variation is not inflated by systematic changes and thus represents better the thresholds approximation for this group of subjects (please see Figure 2).

In this sense, an absolute increase of 10% has been regarded as the smallest worthwhile increase in the frequency of occurrence of anything, regardless of the initial frequency, based partly on converting a frequency difference to a correlation, then applying Cohen's interpretation of magnitude of correlations<sup>2, 16</sup>. If this is so, the smallest important enhancement for recreationally trained subjects is ~0.5% ( $0.3 \times$  typical variation). The thresholds for moderate, large, very large and extremely large are those increasing the chances of winning in 30, 50, 70 and 90%, corresponding respectively to 0.9, 1.6, 2.5 and  $4.0 \times$  the typical error (see Table 2). For example, if a particular treatment results in an enhancement of 1.3% on mean power (~0.9 of the typical error), this strictly means that the chances of the "treatment subject" winning against himself as the "control subject" increased by ~30%. Considering the chances of 50% (with no treatment), now it turns out to be 65%, i.e. about 6-7 times in ten, which can be considered as a moderate effect. These estimates are useful for interpreting



the magnitude of changes in performance, but only when the performance measure is average power. Furthermore, these values are not thresholds for changes in performance times in simulated or actual events, because the relationship between changes in power in laboratory tests and performance in the field are not known<sup>1,3</sup>, and other sources of variation may interfere with performance. However, laboratory-based tests usually seek to evaluate the effects of a particular intervention in a controlled environment set to exclude sources of unwanted variation, which may negatively affect the results by increasing noise.

As can be seen in Table 2, while presenting quite satisfactory reproducibility, researchers using the 1-min sprint cycling as a criterion measure may still experience problems in detecting an effect corresponding to the smallest worthwhile improvement. The low 'signal-to-noise' ratio, i.e. the change in performance divided by the error of measurement<sup>17</sup>, requires a large number of subjects for track small changes in performance, which is an obvious limitation of the test. Nevertheless, when the availability of the participants is a restrictive factor, sensitivity can be improved by increasing the number of trials in each condition<sup>18</sup>. This would reduce noise by a factor corresponding to  $1/\sqrt{n}$ , where  $n$  is the number of trials<sup>12</sup>. Larger effects requires a small number of subjects to be detected; however, a minimum of about 10 subjects is recommended, in order to ensure that they are representative of a wider population.

## CONCLUSION

Long sprint cycling was found to be reliable for recreational cyclists. Based on the present results, the smallest worthwhile improvement in a 1-min cycling performance is near to 0.5%. The estimated thresholds for moderate, large, very large and extremely large effects represent additional benchmarks for assess the magnitude of effects for mean power in repeated measures designs.

## REFERENCES

1. Paton CD, Hopkins WG. Variation in performance of elite cyclists from race to race. *Eur J Sport Sci* 2006;6(1):25-31.
2. Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. *Med Sci Sports Exerc* 1999;31(3):472-85.
3. Currell K, Jeukendrup AE. Validity, reliability and sensitivity of measures of sporting performance. *Sports Med* 2008;38(4):297-316.
4. Siegler JC, McNaughton LR, Midgley AW, Keatley S, Hillman A. Metabolic alkalosis, recovery and sprint performance. *Int J Sports Med* 2010;31(11):797-802.
5. Aisbett B, Le Rossignol P, Sparrow WA. The influence of pacing during 6-minute supra-maximal cycle ergometer performance. *J Sci Med Sport* 2003;6(2):187-98.
6. Hopkins WG. How to interpret changes in an athletic performance test. *Sportscience* 2004;8(1):1-7.
7. Wittekind A, Beneke R. Metabolic and performance effects of warm-up intensity on sprint cycling. *Scand J Med Sci Sports* 2011;21(6):e201-207.
8. De Pauw K, Roelands B, Cheung SS, Geus B, Rietjens G, Meeusen R. Guidelines to classify subject groups in sport-science research. *Int J Sports Physiol Perform* 2013;8(2):111-22.

9. Cruz RS, de Aguiar RA, Turnes T, Salvador AF, Caputo F. Effects of ischemic preconditioning on short-duration cycling performance. *Appl Physiol Nutr Metab* 2016;41(8):825-31.
10. Wittekind AL, Micklewright D, Beneke R. Teleoanticipation in all-out short-duration cycling. *Br J Sports Med* 2011;45(2):114-9.
11. Hopkins WG. Precision of measurement. *A New View of Statistics*. 2011.
12. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000;30(1):1-15.
13. Hopkins WG. A spreadsheet for combining outcomes from several subject groups. *Sportscience* 2006;10(1):51-3.
14. Hopkins WG. Estimating sample size for magnitude-based inferences. *Sportscience* 2006;10(1):63-70.
15. Wiles JD, Coleman D, Tegerdine M, Swaine IL. The effects of caffeine ingestion on performance time, speed and power during a laboratory-based 1 km cycling time-trial. *J Sports Sci* 2006;24(11):1165-71.
16. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum; 1988.
17. Amann M, Hopkins WG, Marcora SM. Similar sensitivity of time to exhaustion and time-trial time to changes in endurance. *Med Sci Sports Exerc*. 2008;40(3):574-578.
18. Paton CD, Hopkins WG. Tests of cycling performance. *Sports Med* 2001;31(7):489-96.

**CORRESPONDING AUTHOR**

Rogério Santos de Oliveira Cruz  
Universidade do Estado de Santa Catarina  
Laboratório de Pesquisas em Desempenho Humano  
Rua Pascoal Simone, 358,  
Coqueiros, Florianópolis, Brasil  
CEP: 88080-350.  
E-mail: cruz.rso@gmail.com