J Appl Physiol 131: 1496–1504, 2021.
First published September 30, 2021; doi:10.1152/japplphysiol.00342.2021

american physiological society

JOURNAL OF APPLIED PHYSIOLOGY

INNOVATIVE METHODOLOGY

Measurement error of self-paced exercise performance in athletic women is not affected by ovulatory status or ambient environment

Huixin Zheng,[1] Claire E. Badenhorst,[2] Tze-Huan Lei,[3] Ahmad Munir Che Muhamed,[4] Yi-Hung Liao,[5] Tatsuro Amano,[6] Naoto Fujii,[7] Takeshi Nishiyasu,[7] Narihiko Kondo,[8] and Toby Mündel[1]

[1]*School of Sport Exercise and Nutrition, Massey University, Palmerston North, New Zealand;* [2]*School of Sport Exercise and Nutrition, Massey University, Auckland, New Zealand;* [3]*College of Physical Education, Hubei Normal University, Huangshi, China;* [4]*Advanced Medical and Dental Institute, Universiti Sains Malaysia, Kepala Batas, Pulau Pinang, Malaysia;* [5]*Department of Exercise and Health Science, National Taipei University of Nursing and Health Sciences, Taipei, Taiwan;* [6]*Faculty of Education, Niigata University, Niigata, Japan;* [7]*Faculty of Health and Sport Sciences, University of Tsukuba, Tsukuba, Japan; and* [8]*Laboratory for Applied Human Physiology, Graduate School of Human Development and Environment, Kobe University, Kobe, Japan*

Abstract

Measurement error(s) of exercise tests for women are severely lacking in the literature. The purpose of this investigation was to *1*) determine whether ovulatory status or ambient environment were moderating variables when completing a 30-min self-paced work trial and *2*) provide test-retest norms specific to athletic women. A retrospective analysis of three heat stress studies was completed using 33 female participants (31 ± 9 yr, 54 ± 10 mL·min^{-1}·kg^{-1}) that yielded 130 separate trials. Participants were classified as ovulatory (n = 19), anovulatory (n = 4), and oral contraceptive pill users (n = 10). Participants completed trials ∼2 wk apart in their (quasi-) early follicular and midluteal phases in two of moderate (1.3 ± 0.1 kPa, 20.5 ± 0.5°C, 18 trials), warm-dry (2.2 ± 0.2 kPa, 34.1 ± 0.2°C, 46 trials), or warm-humid (3.4 ± 0.1 kPa, 30.2 ± 1.1°C, 66 trials) environments. We quantified reliability using limits of agreement, intraclass correlation coefficient (ICC), standard error of measurement (SEM), and coefficient of variation (CV). Test-retest reliability was high, clinically valid (ICC = 0.90, $P < 0.01$), and acceptable with a mean CV of 4.7%, SEM of 3.8 kJ (2.1 W), and reliable bias of −2.1 kJ (−1.2 W). The various ovulatory status and contrasting ambient conditions had no appreciable effect on reliability. These results indicate that athletic women can perform 30-min self-paced work trials ∼2 wk apart with an acceptable and low variability irrespective of their hormonal status or heat-stressful environments.

NEW & NOTEWORTHY This study highlights that aerobically trained women perform 30-min self-paced work trials ∼2 wk apart with acceptably low variability and their hormonal/ovulatory status and the introduction of greater ambient heat and humidity do not moderate this measurement error.

females; heat stress; hormones; performance; reliability

INTRODUCTION

Although *Title IX* and the National Institutes of Health (NIH) Revitalization Act provided a mechanism to prevent the exclusion of females as research participants and focus, there is still a considerable bias against women in basic and preclinical biomedical research, including physiology (1). Nevertheless, this *Journal* and its parent society can claim, more than most, to be leading by example (2–4). The anatomical and physiological characteristics that distinguish the exercise response in women from men indicate a need for recommendations and norms specific to women for exercise testing (5), yet, these data have not been used to determine if sex-specific exercise prescription is necessary (6). Estrogen and progesterone play important secondary (non-reproductive) roles that, according to a specific hormonal milieu, influence physiological systems differently in women, such as vascular, thermal, and osmotic regulation (7, 8). However, women of reproductive age that exercise regularly are more likely to display anovulatory/luteal phase-deficient cycles (30%–50%; 9, 10), and prevalence of oral contraceptive pill (OCP) use among physically active and athletic women is high (>50%; 11, 12). Therefore, when considering research on physically active females, it would be prudent to "include" rather than "exclude" these cohorts because of their physiology (endocrinology), in addition to eumenorrheic women to make findings as representative and applicable to athletic females as possible.

Exercise performance is a common and important outcome measure used to assess the efficacy of treatment effects, such as training programs and other interventions (nutritional, pharmacological, and physiological, etc.). Due

Correspondence: T. Mündel (T.Mundel@massey.ac.nz).
Submitted 19 May 2021 / Revised 28 September 2021 / Accepted 28 September 2021

to the high inter and intravariability in menstrual cycle status and the potential confounding influence on exercise performance, it seems necessary to ascertain if performance effects are due to treatments or due to measurement error (high test or within-/between-subject variability). Another advantage of knowing measurement error is that true differences can be determined without unrealistically large sample sizes, especially if the study design incorporates lifestyle standardization (e.g., diet, time of day, etc.) and a within-subject design. Despite their use in providing mechanistic data during a physiological steady state, constant power tests display inferior reliability and lower ecological validity than constant work or duration tests (13, 14). When considering aerobic tests (i.e., >20 min duration), only one study has previously determined the reliability of a protocol in females. Bishop (15) had 20 female cyclists and triathletes complete two 60-min cycling work trials separated by a week and reported a coefficient of variation (CV) of 2.7%, standard error of measurement (SEM) of 3.4 W, and intraclass correlation coefficient (ICC) of 0.97 for average power output. However, details about the participants' ovulatory/menstrual/hormonal status were not provided.

Given the arguments above, there appears to be a lack of literature describing typical variance of the most commonly utilized exercise testing used for assessing the female physiological response: laboratory cycle ergometry of >20-min duration. We have previously reported that the (quasi-) menstrual phase did not influence the performance of a 30-min work trial, whereas the ambient profile (increased temperature and humidity) reduced work performed by 3%–5% (16–18). Considering that the ~2 wk separating trials between (quasi-) menstrual phases did not affect exercise performance, we used this (within-environment) design to determine test-retest reliability in a homogenous sample of aerobically trained women. This type of retrospective analysis has previously been used to good effect (19). The primary purpose of the current study was to determine whether ovulatory/hormonal status or ambient profile were moderating variables for the measurement error of a 30-min self-paced work trial, whereas the secondary purpose was to add exercise performance norms specific to athletic women to the literature. Given that our previous studies (16–18) observed no (quasi-) menstrual phase-by-ambient profile interaction effect for work completed or mean power output, we hypothesized that measurement error would be unaffected by these factors.

METHODS

Ethical Approval

All previous studies (16–18) had received approval by the Massey University Human Ethics Committee (Southern A) and were performed in accordance with the latest revision of the Declaration of Helsinki, except for registration in a database. Informed, written consent was obtained from all the participants before participation.

Participants

Thirty-three aerobically trained females participated in this study that yielded 130 separate trials, with physical characteristics displayed in Table 1. All were healthy, nonsmokers, and free from cardiovascular, metabolic, neurological, and respiratory diseases and were not taking any regular medication apart from those using the OCP. Some of the data herein have been reported previously in separate studies (16–18). All eumenorrheic females self-reported a regular menstrual cycle 21–35 days in length (≥3 mo) with no use of hormonal contraception (≥6 mo). All females taking OCP were taking a monophasic combination OCP (≥1 yr) with experimental visits completed during the 3 wk of active pill use (see Ref. 17 for further details).

Experimental Overview

Data collection was conducted excluding the southern hemisphere summer (March–November) where the average daily temperature did not exceed 22°C, nor had participants spent any time in a warmer climate for at least 1 mo before the study. All the participants attended the laboratory on six occasions: *1*) preliminary submaximal and maximal aerobic

Table 1. *Participant characteristics for ovulatory, anovulatory, and oral contraceptive pill groups*

Characteristic	OVU	ANO	OCP	Mean	*P* Value
n	19	4	10	33	
Age, yr	34 (9)	36 (8)	25 (5)*	31 (9)	0.02
Mass, kg	63 (6)	65 (3)	68 (10)	65 (7)	0.28
$A_{D,}$ m^2	1.70 (0.11)	1.69 (0.03)	1.76 (0.13)	1.72 (0.11)	0.45
A_D: mass	0.027 (0.001)	0.026 (0.001)	0.026 (0.002)	0.027 (0.001)	0.30
% fat	23 (5)	22 (6)	24 (5)	23 (5)	0.70
$\dot{V}O_{2max}$, $L \cdot min^{-1}$	3.3 (0.6)	3.8 (1.0)	3.7 (0.5)	3.5 (0.6)	0.11
$\dot{V}O_{2max}$, $mL \cdot min^{-1} \cdot kg^{-1}$	52 (9)	58 (15)	55 (9)	54 (10)	0.40
Wmax, W	270 (40)	292 (39)	283 (29)	276 (37)	0.46
Training history, yr	7.3 (3.3)	9.3 (5.3)	3.7 (2.5)*	6.4 (3.8)	0.01
Progesterone, $ng \cdot mL^{-1}$					
Follicular	0.6 (0.4)*	0.2 (0.1)	0.1 (0.1)		<0.01
Luteal	16.1 (12.2)*	1.2 (1.5)	0.2 (0.1)		<0.01
Estrogen, $pg \cdot mL^{-1}$					
Follicular	63.6 (53.1)	42.5 (15.0)	18.3 (23.0)†		<0.01
Luteal	105.2 (77.2)	90.6 (71.2)	20.5 (28.7)*		<0.01

Values are means (SD); *n*, number of women per group. *Significantly different to both other groups. †Significantly different to OVU. $A_{D,}$ Du Bois body surface area; ANO, anovulatory; OCP, oral contraceptive pill; OVU, ovulatory; $\dot{V}O_{2max}$, maximal rate of O_2 consumption; Wmax, peak aerobic power.

capacity test, *2*) experimental familiarization, *3–6*) experimental trials. For a diagrammatic representation of the experimental overview, see Fig. 1. The four experimental trials were a full crossover of (quasi-) menstrual phase [early follicular (EF, 65 trials) and midluteal (ML, 65 trials)] and ambient profile [moderate (MOD, 19 trials), warm-humid (HUM, 65 trials), or warm-dry (DRY, 46 trials)]. Grouping for the ambient conditions was based on vapor pressure, such that the following characterized each environment: MOD [1.3 ± 0.1 kPa, 20.5 ± 0.5°C, 53.1 ± 5.5 relative humidity (rh)], DRY (2.2 ± 0.2 kPa, 34.1 ± 0.2°C, 41.4 ± 3.4 rh), and HUM (3.4 ± 0.1 kPa, 30.2 ± 1.1°C, 80.0 ± 3.7 rh). The order of the trials was randomized and counterbalanced except the order of the ambient profile was consistent in different (quasi-) phases within participants. Experimental trials were conducted at the same time of day (±1 h) and following >24 h of dietary and exercise control. Each trial consisted of either 12 (92 trials) or 20 (38 trials) min of fixed-intensity preload immediately followed by a 30-min self-paced work trial where only percentage of time elapsed (every 20% or 6 min) was provided to the participant. All exercise was performed on an electronically braked cycle ergometer (Lode Excalibur, Groningen, The Netherlands), with handlebars, seat height, and pedal preference standardized according to individual preference. The typical timeline for a participant to complete this study resulted in preliminary testing and familiarization separated by 3–7 days during the (quasi-) follicular phase, with half of the participants starting their experimental trials the following (quasi-) luteal phase (14 days later) and the other half the following (quasi-) follicular phase (28 days later), with within-phase experimental trials differing by ambient profile separated by 3 days.

Preliminary Testing and Familiarization

All preliminary testing was conducted in the (quasi-) EF phase of each participant's menstrual cycle to minimize the potential effects of the menstrual/OCP cycle on their physiological and performance responses during the tests (8). Following anthropometric measurements (height, weight, and body composition), a 24-min steady-state submaximal exercise test was conducted in a temperate laboratory environment (18°C–22°C). The submaximal exercise test consisted of four consecutive 6-min stages with power outputs of 100 W, 125 W, 150 W, and 175 W at a comfortable but constant cadence. Oxygen consumption was measured during the last 2 min of each stage. Following 10 min rest from the submaximal test, a maximal oxygen consumption ($\dot{V}O_{2max}$) test was performed. The initial workload began at 100 W and increased by 25 W every minute, until volitional exhaustion. The exercise "gear" (linear factor, see *Experimental Procedure*) during the self-paced exercise was based on 75% of an individual's $\dot{V}O_{2max}$, which was derived from the linear relationship between the power output and the O_2 consumption during both the steady-state submaximal exercise test and maximal aerobic capacity test. Following at least 24-h rest from the preliminary session, a familiarization trial was conducted to ensure all participants were familiar with the testing procedures and to minimize the learning effect

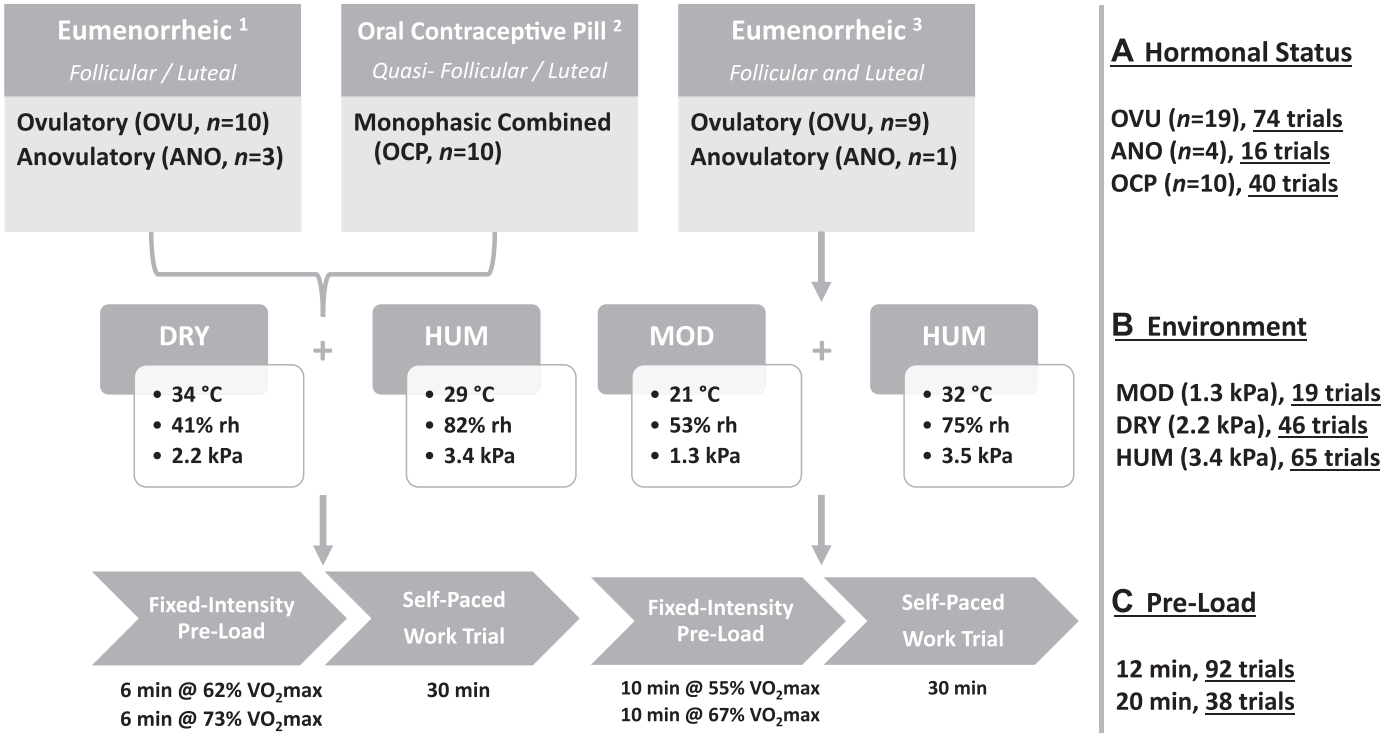

Figure 1. Diagram of experimental overview. Three distinct groups were determined according to their hormonal status (*A*), with each participant performing trials under two distinct ambient environments (*B*) and repeated ~2 wk later according to their (quasi-) follicular and luteal phases (*A*). Each 30-min self-paced work trial was preloaded with either a 12 or 20 min fixed-intensity period (*C*). [1]participants for Lei et al. (16); [2]participants for Lei et al. (17); [3]participants for Zheng et al. (18). DRY, warm-dry; HUM, warm-humid; MOD, moderate; OCP, oral contraceptive pill; $\dot{V}O_{2max}$, maximal oxygen consumption.

during trials. This trial replicated entirely the experimental trials outlined below (see *Experimental Procedure*).

Dietary and Exercise Control

Diet and physical activity during the 48 h before the first experimental trial were recorded, and the participants were instructed to repeat these for the following trials. The day of and before any experimental trial was marked by abstinence from alcohol, exercise, and only habitual caffeine use (as abstinence would confound results from withdrawal effects). This dietary and exercise control minimized variation in pretrial metabolic state. Fluid intake was encouraged to ensure a euhydrated state.

Ovulatory Status and Ambient Conditions

Eumenorrheic females were tested on *days 3–6* (EF) and *18–21* (ML) following the start of menses, whereas OCP females were tested on *days 3–6* and *18–21* following the start of OCP use. Testing for eumenorrheic females was scheduled using the three-step method (20), whereby self-reported menses onset and urinary luteinizing hormone testing (EasyCheck Ovulation Test, Phoenix Medcare Ltd, Auckland, New Zealand) prospectively identified EF and ML, whereas measurement of serum 17β-estradiol and progesterone retrospectively confirmed ML. A progesterone level of >5 $ng{\cdot}mL^{-1}$ is good evidence that ovulation has occurred (10, 21, 22). Therefore, the participants were deemed as ovulatory (OVU, >5 $ng{\cdot}mL^{-1}$) or anovulatory (ANO, <5 $ng{\cdot}mL^{-1}$).

Experimental Procedure

These four trials were conducted in the same environmental chamber with a fan-generated airflow of 19 $km{\cdot}h^{-1}$. Upon their arrival at the laboratory, the participants voided and a blood sample was obtained from an antecubital vein after participants had rested seated for 15 min. Participants entered the environmental chamber wearing only cycling shorts and top, shoes, and socks. Participants rested seated on the ergometer for 20 min before completing either *1*) 6 min of cycling at each of 125 and 150 W (62 ± 9 and 73 ± 10% $\dot{V}O_{2max}$, respectively) or *2*) 10 min of cycling at each of 100 and 125 W (55 ± 8 and 67 ± 9% $\dot{V}O_{2max}$, respectively). Immediately on completion of the second fixed-intensity bout, the ergometer was set to linear mode based on the formula of Jeukendrup et al. (23), where participants were instructed to perform as much work as possible over 30 min. During this 30-min self-paced period, work completed (kJ) was recorded every 6 min and tap water at 20°C was provided to drink ad libitum throughout to minimize dehydration. Total work completed (kJ) was used as the criterion measure for reliability metrics (see *Statistical Analysis*).

Measurements

For interested readers, other physiological measurements (i.e., thermoregulatory, cardiovascular, and inflammatory) were performed during these trials of which the results can be found in our separate studies (16–18).

Anthropometric.

Participant height and weight were measured using a stadiometer (Seca, Germany; accurate to 0.1 cm) and scale (Jadever, Taiwan; accurate to 0.01 kg), from which surface area was estimated (24). Body composition was measured using multifrequency bioelectrical impedance analysis (InBody 230, Korea) using a standard procedure (25).

Respiratory.

Expired respiratory gases were collected from a mixing chamber and analyzed for O_2 consumption using an online, breath-by-breath system (VacuMed Vista, Turbofit, Ventura, CA) using a 30-s average. This system was calibrated before each trial using a zero and β-standard gas concentrations and volume (VacuMed 3 L Calibration Syringe).

Hormones.

Venous blood was collected by venipuncture into a vacutainer (Becton Dickinson, Oxford, UK) containing clot activator, and once clotted (>30 min), the whole blood was centrifuged at 4°C and 805 *g* for 15 min and aliquots of serum were transferred into Eppendorf tubes (Genuine Axygen Quality) and stored at −80°C till further analysis. For further detail, please see our previous studies (16–18).

Statistical Analysis

All statistical analyses were performed with SPSS software for Windows (IBM SPSS Statistics 20, NY). Descriptive values were obtained and reported as means and standard deviation (SD) or using 95% confidence interval (CI) unless stated otherwise. Homogeneity of variance was examined by Levene's test, whereas the normality of the data was examined by the Shapiro–Wilk test, with no significant effects. Participant characteristics were analyzed using a one-way analysis of variance. To assess test-retest reliability in work trial performance (work completed in kJ), several commonly reported measures were calculated (26). Limits of agreement (LoA; 27) are reported as bias ± 1.96 SD. Giavarina (28) proposed that if the line of equality ($x = 0$) lies within the 95% CI (±1.96 SE) of the mean of the differences, the bias is not significant and the measurement is reliable. The ICC was calculated based on an absolute agreement, two-way mixed-effects model with high and clinically valid reliability denoted as >0.9 (29, 30). The SEM (also known as typical error) was calculated as $SD\sqrt{(1 - ICC)}$ (31). The within-subject CV was calculated as the SD divided by the mean of two repeated trials performed under the same ambient conditions, then multiplied by 100%; we used <5% as acceptable reliability (14). Finally, Pearson's correlation coefficient was used to examine the direction and strength of relationships between the independent (participant characteristics) and dependent (work completed in kJ) variables; this was deemed appropriate as data were continuous, related pairs, normally distributed, linear, and homoscedastic with minimal outliers. Statistical significance was set at $P < 0.05$.

RESULTS

Although the OCP group was significantly younger with less training history, all three groups were not different on most physical characteristics (Table 1). A higher concentration of progesterone characterized the OVU group, whereas estrogen was suppressed in the OCP group.

With regard to repeated performance of the work trials, individual results can be seen as Brinley and Bland–Altman plots (Fig. 2). Results were presented as homoscedastic, although two clear outliers were identified a posteriori using Tukey's method (interquartile range × 1.5); where data were identified as outliers, these points (i.e., reliability between two trials) were removed. Therefore, as all outliers belonged to the same two participants, all of their data (four trials) were removed. Overall bias was −2.1 ± 54.6 kJ, with this bias seemingly not significant and the measurement reliable (28).

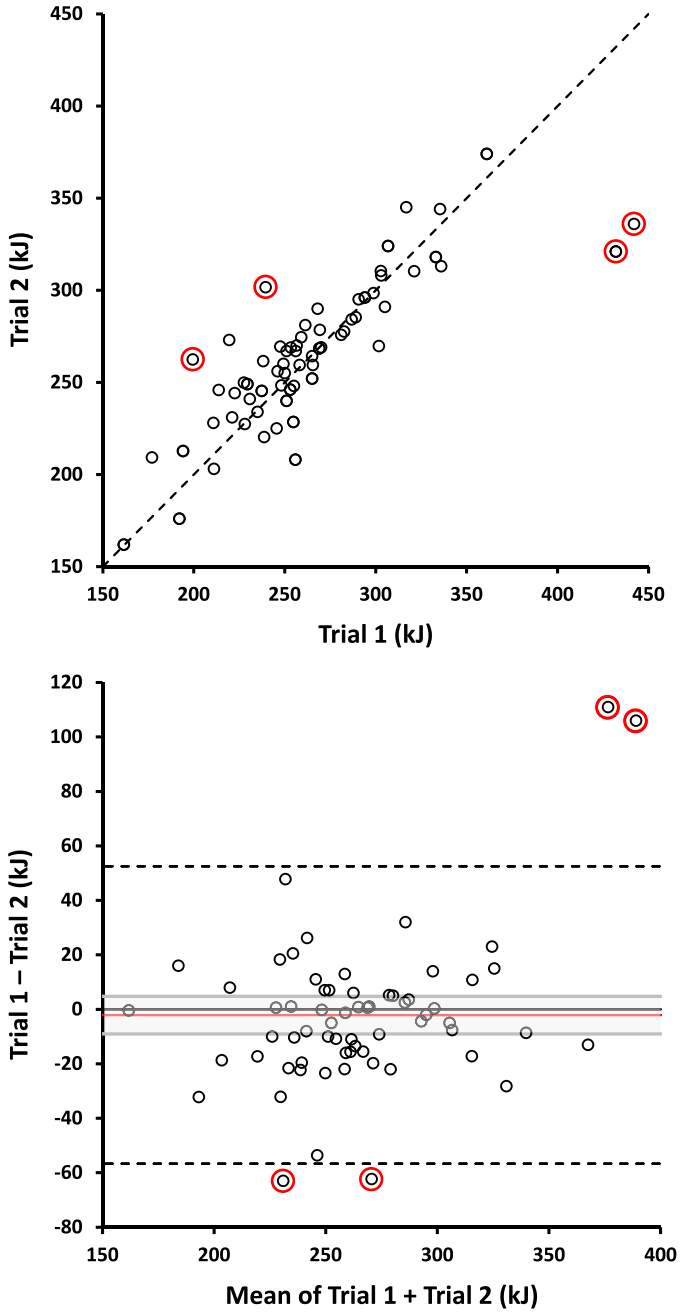

Figure 2. Brinley (*top*) and Bland–Altman (*bottom*) plots displaying individual work trial results. Two identified outliers circled red. For Bland–Altman plot, solid black line, line of equality; red line, bias; dashed black lines, limits of agreement; gray lines, confidence limits; shaded area, confidence limits for mean.

Table 2 displays results for the different ways of assessing reliability between repeated work trials, with these results displayed graphically in Figs. 3 and 4. Overall, reliability was high, clinically valid (ICC = 0.90, $P < 0.01$), and acceptable with a mean CV of 4.7%, SEM of 3.8 kJ (2.1 W), and reliable bias of −2.1 kJ (−1.2 W). The different duration of preload and varying ambient conditions had no appreciable effect on reliability, whereas it could be argued that hormonal status (ANO ≠ OVU/OCP) affected reliability due to *1*) the ICC ($P = 0.05$) not being different to 0, *2*) having the largest CI for all metrics, and *3*) the largest SEM (7.7 kJ or 4.3 W). However, on closer inspection, this is likely a reflection of sample size as demonstrated best by the removal of the two identified outliers (circled red, Fig. 2). Following the removal of the two outliers, the change in values was greatest for ANO for ICC [OVU: 0.96 (0.91–0.98), ANO: 0.98 (0.88–1.00), Overall: 0.95 (0.91–0.97), all $P < 0.01$], % CV [OVU: 4.1 (2.7–5.4), ANO: 1.7 (0.2–3.1), Overall: 3.8 (2.9–4.7)], and SEM (OVU: 2.0 kJ, ANO: 0.7 kJ, Overall: 2.1 kJ).

To determine the potential factors contributing to the variance in work trial performance, bivariate correlations were performed between the within-participant % CV and physical characteristics in Table 1. In absolute terms, only peak aerobic power (Wmax, W) correlated with % CV ($r = 0.37$, $P < 0.01$); all other correlations $r < 0.20$ and $P > 0.10$.

DISCUSSION

The important results from this investigation were that: *1*) hormonal/ovulatory status and the introduction of greater ambient heat and humidity did not moderate measurement error of 30-min self-paced work trials that were completed ~2 wk apart and *2*) aerobically trained women repeatedly performed this test with acceptably low variability.

Previously, Bishop (15) observed a CV of 2.7%, SEM of 3.4 W, and ICC of 0.97 for average power output in female cyclists and triathletes completing two 60-min cycling work trials separated by a week. Our results are comparable, observing a CV of 4.7%, SEM of 2.1 W, and ICC of 0.90. Taken together, the results of these two studies support the use of a self-paced work trial for the assessment of aerobic performance (or behavior) to determine intervention success in trained women and indicate little difference to trained males (CV <5% as acceptable reliability; 14) despite early concerns (13). Moreover, it has been speculated that increased ambient temperatures could introduce further variance to self-paced (constant work and duration) trials (19), although the evidence from studies in trained men is to the contrary, i.e., a CV ≤ 3.6% irrespective of ambient temperature and humidity (23, 32, 33). The current results (Table 2, Figs. 3 and 4) support the latter and corroborate these findings to trained women, thereby refuting the original supposition (19).

Recent meta-analyses of the literature concluded trivial effects (Cohen's $d < 0.2$) of OCP use and menstrual cycle phase on exercise performance (34, 35), indicating little meaningful effect from the variance in endogenous or exogenous ovarian hormone concentrations. The current results

Table 2. *Measures of reliability for work completed (kJ) during the 30-min self-paced cycling work trial*

	Work (kJ, 95% CI)	ICC (95% CI)	CV (%, 95% CI)	SEM, kJ
Hormonal Status				
Ovulatory	265 (253–277)	0.91* (0.82–0.95)	4.9 (3.1–6.7)	4.1
Anovulatory	265 (249–281)	0.72 (–0.15 to 0.94)	5.6 (–0.6 to 11.9)	7.7
OCP	265 (256–275)	0.91* (0.77–0.97)	4.0 (2.7–5.2)	3.1
Environment				
Moderate	277 (250–303)	0.86* (0.35–0.97)	5.9 (0.7–11.0)	6.7
Dry	269 (259–279)	0.91* (0.78–0.96)	4.1 (2.3–5.8)	3.2
Humid	260 (248–271)	0.91* (0.81–0.95)	4.9 (3.1–6.6)	3.8
Preload				
12 min	260 (253–268)	0.91* (0.82–0.95)	4.5 (3.2–5.8)	3.4
20 min	277 (258–296)	0.88* (0.70–0.96)	5.3 (2.3–8.2)	5.6
Overall	*265 (257–273)*	*0.90* (0.83–0.94)*	*4.7 (3.5–5.9)*	*3.8*

*Significant at $P < 0.01$ following analysis of variance. CI, confidence interval; CV, coefficient of variation; ICC, intraclass correlation coefficient; OCP, oral contraceptive pill; SEM: standard error of measurement.

are the first to determine performance reproducibility in relation to trained but hormonally distinct women, with an important finding that women often excluded a posteriori from physiological investigations due to subtle menstrual disturbances, display similar reliability to eumenorrheic athletes (Table 2, Figs. 3 and 4). This is likely reflective of or consequent to similar body composition, functional capacity, and training history (Table 1), as supported by these characteristics not being correlated with percent CV. However, Wmax did correlate positively with % CV, and the two identified outliers (Fig. 2) were two of the "best" performers—both in terms of Wmax (95th percentile) and race performance/results: one a former professional and continental road cycling champion, whereas the other a national masters road cycling champion.

Total measurement error is composed of both systematic bias (e.g., learning or fatigue effects) and random error (e.g., biological or mechanical variation), with both components ideally quantified (26). The most common methods for assessing relative reliability (*r* and ICC) are highly influenced by the range of values in the sample and cannot by themselves assess systematic bias, therefore, should not be used to extrapolate results to new individuals or compare between different measurement tools (26). The most common methods for assessing absolute reliability overcome some of these issues and are expressed in the actual unit of the measurement (SEM) or dimensionless (CV), although represent (only) ~68% of the error present for an average individual (26). The LoA quantifies systematic bias (−2.1 kJ or −1.2 W) and random error (±54.6 kJ or 30.3 W), such that for any new female athlete, the difference between her two work trials should lie within these limits with a ~95% probability. It is also possible to determine a signal-to-noise ratio (sensitivity index) knowing total measurement error, such that one can be confident of the true effect of an intervention on performance. See APPENDIX for a worked example.

Considerations

The observations herein are valid only for the current sample(s), protocol(s), and condition(s). Although the time between trials (~2 wk) in the current study is both longer (weekly trials for males) and shorter (4 wk +, training interventions) than might be required, Hopkins et al. (13) concluded that time between trials is likely to have a smaller effect than the type of test and measure used, athletic status, and duration of test. We were unable to distinguish between anovulatory or luteal phase-deficient cycles as detection of a urinary luteinizing hormone surge (alone) cannot confirm luteal phase sufficiency without a reduction in estrogen concentrations (36, 37), which warrants further investigation. It is noteworthy that (sub-) sample size appeared to affect our results, although impossible to improve/increase in such a retrospective analysis. The two smallest data samples (<40, Fig. 1), the anovulatory group and moderate ambient condition, displayed considerably lower reliability than the other two groups/conditions (Table 2). However, as demonstrated in the RESULTS, omission of only one of the identified outliers improved reliability disproportionately compared with the larger data set. This is a well-known phenomenon of small data sets as they are more sensitive to heterogeneity and why only larger samples (>40) are recommended to be examined by limits of agreement (26). Thus, we caution interpretation of the current results, especially for these two subsamples, with further confirmatory research required with a priori data samples >40.

Perspectives and significance

Global warming and urbanization present a current and increasing threat to human health and performance/productivity, with nearly one-third of the global population regularly exposed to extreme heat events and a pertinent risk factor being increased physical activity for sport, exercise, or occupation (38, 39). Concurrently, the number of women working in physically demanding occupations (e.g., mining, logging, construction, firefighting, and military etc.) continues to increase (40), alongside the rising number of sports open to, events for, and number of competitive female athletes (41). Thus, mitigation strategies (or ergogenic aids) for active women encountering this combined heat load (metabolic and environmental) are warranted. The current study should provide the impetus for being able to correctly identify the true effects of, for example, exercise training, heat acclimation, hydration, cooling, and dietary interventions (42–45) in the athletic female population.

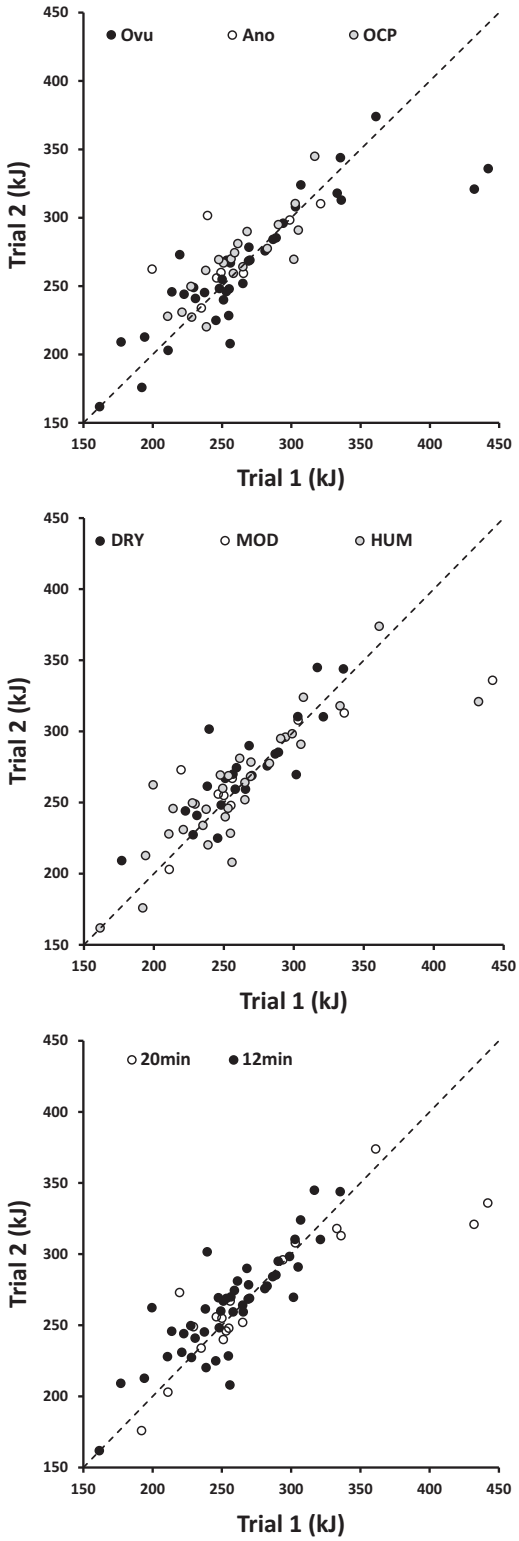

Figure 3. Brinley plots displaying individual work trial results and grouped by hormonal status (*top*, OVU, ovulatory; ANO, anovulatory; OCP, oral contraceptive pill), ambient condition (*middle*, MOD, moderate; DRY, warm-dry; HUM, warm-humid), and preload duration (*bottom*).

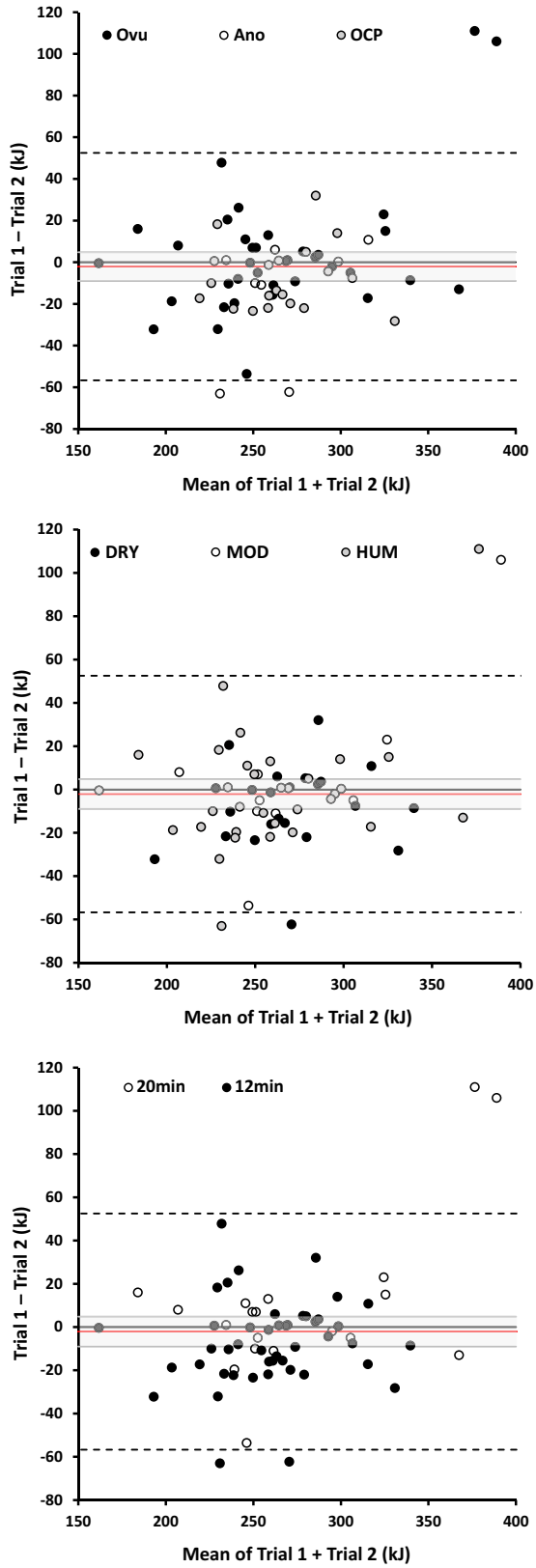

Figure 4. Bland–Altman plots displaying individual work trial results and grouped by hormonal status (*top*, OVU, ovulatory; ANO, anovulatory; OCP, oral contraceptive pill), ambient condition (*middle*, MOD, moderate; DRY, warm-dry; HUM, warm-humid), and preload duration (*bottom*).

APPENDIX

Calculating a Signal-to-Noise Ratio (Sensitivity Index)

Currell and Jeukendrup (14) proposed a quantitative measure of sensitivity expressed as the ratio between an intervention "signal" and measurement error "noise," whereby a higher value indicates greater protocol sensitivity. A sensitivity index ≤ 1 infers that the test completed under those conditions and with that sample is not sufficiently sensitive/reliable or that the true effect of the intervention is small/negligible. The first step is to identify the change in performance and within-subject CV. From Table 2, we can see that the CV for OCP is 4.0%, whereas the effect of a reduction in ambient thermal stress (lower vapor pressure) results in a change in performance of 5.5% in this cohort (17). Therefore, relating the signal (5.5%) to the noise (4.0%) means solving the equation:

sensitivity index = signal/noise

or

$1.4 = 5.5/4.0$.

GRANTS

This study was supported by the New Zealand-Japan Joint Research Project Programme, under Catalyst: seeding funding from Royal Society Te Apārangi (T.M.).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

H.Z., T-H.L., and T.M. conceived and designed research; H.Z., T-H.L., and T.M. performed experiments; H.Z., C.E.B., T-H.L., and T.M. analyzed data; H.Z., C.E.B., T-H.L., and T.M. interpreted results of experiments; H.Z., T-H.L., and T.M. prepared figures; H.Z., C.E.B., T-H.L., and T.M. drafted manuscript; H.Z., C.E.B., T-H.L., A.M.C.M., Y-H.L., T.A., N.F., T.N., N.K., and T.M. edited and revised manuscript; H.Z., T.M., C.E.B., T.L., A.C., Y.L., T.A., N.F., T.N. and N.K. approved final version of manuscript.

REFERENCES

1. **Beery AK**, **Zucker I.** Sex bias in neuroscience and biomedical research. *Neurosci Biobehav Rev* 35: 565–572, 2011. doi:10.1016/j.neubiorev.2010.07.002.
2. **Stachenfeld NS.** Women leading in the environmental and exercise physiology section of the American Physiological Society: better late than never. *J Appl Physiol (1985)* 127: 893, 2019. doi:10.1152/japplphysiol.00295.2019.
3. **Wenner MM**, **Stachenfeld NS.** Point: Investigators should control for menstrual cycle phase when performing studies of vascular control that include women. *J Appl Physiol (1985)* 129: 1114–1116, 2020. doi:10.1152/japplphysiol.00443.2020.
4. **Stanhewicz AE**, **Wong BJ.** Counterpoint: investigators should not control for menstrual cycle phase when performing studies of vascular control that include women. *J Appl Physiol (1985)* 129: 1117–1119, 2020. doi:10.1152/japplphysiol.00427.2020.
5. **Charkoudian N**, **Joyner MJ.** Physiologic considerations for exercise performance in women. *Clin Chest Med* 25: 247–255, 2004. doi:10.1016/j.ccm.2004.01.001.
6. **Liguori G**, **Feito Y**, **Fountaine C**, **Roy B**; **American College of Sports Medicine.** *ACSM's Guidelines for Exercise Testing and Prescription*. Philadelphia: Wolters Kluwer, 2021.
7. **Charkoudian N**, **Stachenfeld NS.** Reproductive hormone influences on thermoregulation in women. *Compr Physiol* 4: 793–804, 2014. doi:10.1002/cphy.c130029.
8. **Sims ST**, **Heather AK.** Myths and methodologies: reducing scientific design ambiguity in studies comparing sexes and/or menstrual cycle phases. *Exp Physiol* 103: 1309–1317, 2018. doi:10.1113/EP086797.
9. **De Souza MJ**, **Toombs RJ**, **Scheid JL**, **O'Donnell E**, **West SL**, **Williams NI.** High prevalence of subtle and severe menstrual disturbances in exercising women: confirmation using daily hormone measures. *Hum Reprod* 25: 491–503, 2010. doi:10.1093/humrep/dep411.
10. **Schaumberg MA**, **Jenkins DG**, **Janse de Jonge XAK**, **Emmerton LM**, **Skinner TL.** Three-step method for menstrual and oral contraceptive cycle verification. *J Sci Med Sport* 20: 965–969, 2017. doi:10.1016/j.jsams.2016.08.013.
11. **Rechichi C**, **Dawson B**, **Goodman C.** Athletic performance and the oral contraceptive. *Int J Sports Physiol Perform* 4: 151–162, 2009. doi:10.1123/ijspp.4.2.151.
12. **Martin D**, **Sale C**, **Cooper SB**, **Elliott-Sale KJ.** Period prevalence and perceived side effects of hormonal contraceptive use and the menstrual cycle in elite athletes. *Int J Sports Physiol Perform* 13: 926–932, 2018. doi:10.1123/ijspp.2017-0330.
13. **Hopkins WG**, **Schabort EJ**, **Hawley JA.** Reliability of power in physical performance tests. *Sports Med* 31: 211–234, 2001. doi:10.2165/00007256-200131030-00005.
14. **Currell K**, **Jeukendrup AE.** Validity, reliability and sensitivity of measures of sporting performance. *Sports Med* 38: 297–316, 2008. doi:10.2165/00007256-200838040-00003.
15. **Bishop D.** Reliability of a 1-h endurance performance test in trained female cyclists. *Med Sci Sports Exerc* 29: 554–559, 1997. doi:10.1097/00005768-199704000-00019.
16. **Lei TH**, **Stannard SR**, **Perry BG**, **Schlader ZJ**, **Cotter JD**, **Mündel T.** Influence of menstrual phase and arid vs. humid heat stress on autonomic and behavioural thermoregulation during exercise in trained but unacclimated women. *J Physiol* 595: 2823–2837, 2017. doi:10.1113/JP273176.
17. **Lei TH**, **Cotter JD**, **Schlader ZJ**, **Stannard SR**, **Perry BG**, **Barnes MJ**, **Mündel T.** On exercise thermoregulation in females: interaction of endogenous and exogenous ovarian hormones. *J Physiol* 597: 71–88, 2019. doi:10.1113/JP276233.
18. **Zheng H**, **Badenhorst CE**, **Lei TH**, **Liao YH**, **Che Muhamed AM**, **Fujii N**, **Kondo N**, **Mundel T.** Menstrual phase and ambient temperature do not influence iron regulation in the acute exercise period. *Am J Physiol Regul Integr Comp Physiol* 320: R780–R790, 2021. doi:10.1152/ajpregu.00014.2021.
19. **Salgado RM**, **Caldwell AR**, **Coffman KE**, **Cheuvront SN**, **Kenefick RW.** Endurance test selection optimized via sample size predictions. *J Appl Physiol (1985)* 129: 467–473, 2020. doi:10.1152/japplphysiol.00408.2020.
20. **Allen AM**, **McRae-Clark AL**, **Carlson S**, **Saladin ME**, **Gray KM**, **Wetherington CL**, **McKee SA**, **Allen SS.** Determining menstrual phase in human biobehavioral research: a review with recommendations. *Exp Clin Psychopharmacol* 24: 1–11, 2016. doi:10.1037/pha0000057.
21. **Leiva R**, **Bouchard T**, **Boehringer H**, **Abulla S**, **Ecochard R.** Random serum progesterone threshold to confirm ovulation. *Steroids* 101: 125–129, 2015. doi:10.1016/j.steroids.2015.06.013.
22. **Janse DE Jonge X**, **Thompson B**, **Han A.** Methodological recommendations for menstrual cycle research in sports and exercise. *Med Sci Sports Exerc* 51: 2610–2617, 2019. doi:10.1249/MSS.0000000000002073.
23. **Jeukendrup A**, **Saris WH**, **Brouns F**, **Kester AD.** A new validated endurance performance test. *Med Sci Sports Exerc* 28: 266–270, 1996. doi:10.1097/00005768-199602000-00017.
24. **Du Bois D**, **Du Bois EF.** A formula to estimate approximate surface area if height and weight be known. *Arch Intern Med (Chic)* 17: 863–871, 1916. doi:10.1001/archinte.1916.00080130010002.
25. **Kyle UG**, **Bosaeus I**, **De Lorenzo AD**, **Deurenberg P**, **Elia M**, **Gómez JM**, **Heitmann BL**, **Kent-Smith L**, **Melchior J-C**, **Pirlich M**, **Scharfetter H**, **Schols AMWJ**, **Pichard C**; **Composition of the ESPEN Working Group.** Bioelectrical impedance analysis—part I: review of principles and methods. *Clin Nutr* 23: 1226–1243, 2004. doi:10.1016/j.clnu.2004.06.004.

26. **Atkinson G**, **Nevill AM.** Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 26: 217–238, 1998. doi:10.2165/00007256-199826040-00002.
27. **Bland JM**, **Altman DG.** Measuring agreement in method comparison studies. *Stat Methods Med Res* 8: 135–160, 1999. doi:10.1177/096228029900800204.
28. **Giavarina D.** Understanding Bland Altman analysis. *Biochem Med (Zagreb)* 25: 141–151, 2015. doi:10.11613/BM.2015.015.
29. **Portney LG**, **Watkins MP.** *Foundations of Clinical Research: Applications to Practice*. Philadelphia: F.A. Davis, 2020.
30. **Weir JP**, **Vincent WJ.** *Statistics in Kinesiology*. Champaign, IL: Human Kinetics, 2021.
31. **Hopkins WG.** Measures of reliability in sports medicine and science. *Sports Med* 30: 1–15, 2000. doi:10.2165/00007256-200030010-00001.
32. **Marino FE**, **Kay D**, **Cannon J**, **Serwach N**, **Hilder M.** A reproducible and variable intensity cycling performance protocol for warm conditions. *J Sci Med Sport* 5: 95–107, 2002. doi:10.1016/S1440-2440(02)80030-5.
33. **Che Jusoh MR**, **Morton RH**, **Stannard SR**, **Mündel T.** A reliable preloaded cycling time trial for use in conditions of significant thermal stress. *Scand J Med Sci Sports* 25, *Suppl*1: 296–301, 2015. doi:10.1111/sms.12332.
34. **Elliott-Sale KJ**, **McNulty KL**, **Ansdell P**, **Goodall S**, **Hicks KM**, **Thomas K**, **Swinton PA**, **Dolan E.** The effects of oral contraceptives on exercise performance in women: a systematic review and meta-analysis. *Sports Med* 50: 1785–1812, 2020. doi:10.1007/s40279-020-01317-5.
35. **McNulty KL**, **Elliott-Sale KJ**, **Dolan E**, **Swinton PA**, **Ansdell P**, **Goodall S**, **Thomas K**, **Hicks KM.** The effects of menstrual cycle phase on exercise performance in eumenorrheic women: a systematic review and meta-analysis. *Sports Med* 50: 1813–1827, 2020. doi:10.1007/s40279-020-01319-3.
36. **Elliott-Sale KJ**, **Minahan CL**, **de Jonge XAKJ**, **Ackerman KE**, **Sipilä S**, **Constantini NW**, **Lebrun CM**, **Hackney AC.** Methodological considerations for studies in sport and exercise science with women as participants: a working guide for standards of practice for research on women. *Sports Med* 51: 843–861, 2021. doi:10.1007/s40279-021-01435-8.
37. **Scheid JL**, **De Souza MJ.** Menstrual irregularities and energy deficiency in physically active women: the role of ghrelin, PYY and adipocytokines. *Med Sport Sci* 55: 82–102, 2010. doi:10.1159/000321974.
38. **Luber G**, **McGeehin M.** Climate change and extreme heat events. *Am J Prev Med* 35: 429–435, 2008. doi:10.1016/j.amepre.2008.08.021.
39. **Mora C**, **Dousset B**, **Caldwell IR**, **Powell FE**, **Geronimo RC**, **Bielecki CR**, **Counsell CWW**, **Dietrich BS**, **Johnston ET**, **Louis LV**, **Lucas MP**, **McKenzie MM**, **Shea AG**, **Tseng H**, **Giambelluca TW**, **Leon LR**, **Hawkins E**, **Trauernicht C.** Global risk of deadly heat. *Nature Clim Change* 7: 501–506, 2017. doi:10.1038/nclimate3322.
40. **U.S. Department of Labor.** BLS Spotlight on Statistics: Women at Work, 2017 https://www.bls.gov/spotlight/2017/women-at-work/.
41. **International Olympic Committee.** IOC Factsheet - Women in the Olympic Movement, 2020 https://olympics.com/ioc/faq/the-ioc-and-the-olympic-movement-commitment-to-integrity/what-is-the-role-of-women-in-the-olympic-movement.
42. **Grahn DA**, **Cao VH**, **Heller HC.** Heat extraction through the palm of one hand improves aerobic exercise endurance in a hot environment. *J Appl Physiol (1985)* 99: 972–978, 2005. doi:10.1152/japplphysiol.00093.2005.
43. **Szymanski MC**, **Gillum TL**, **Gould LM**, **Morin DS**, **Kuennen MR.** Short-term dietary curcumin supplementation reduces gastrointestinal barrier damage and physiological strain responses during exertional heat stress. *J Appl Physiol (1985)* 124: 330–340, 2018. doi:10.1152/japplphysiol.00515.2017.
44. **Ravanelli N**, **Coombs G**, **Imbeault P**, **Jay O.** Thermoregulatory adaptations with progressive heat acclimation are predominantly evident in uncompensable, but not compensable, conditions. *J Appl Physiol (1985)* 127: 1095–1106, 2019. doi:10.1152/japplphysiol.00220.2019.
45. **Chapman CL**, **Johnson BD**, **Vargas NT**, **Hostler D**, **Parker MD**, **Schlader ZJ.** Both hyperthermia and dehydration during physical work in the heat contribute to the risk of acute kidney injury. *J Appl Physiol (1985)* 128: 715–728, 2020. doi:10.1152/japplphysiol.00787.2019.