J Appl Physiol 129: 467–473, 2020.
First published July 30, 2020; doi:10.1152/japplphysiol.00408.2020.

INNOVATIVE METHODOLOGY

Endurance test selection optimized via sample size predictions

Roy M. Salgado,[1] Aaron R. Caldwell,[1] Kirsten E. Coffman,[1] Samuel N. Cheuvront,[2] and Robert W. Kenefick[1]

[1]*Thermal and Mountain Medicine Division, US Army Research Institute of Environmental Medicine, Natick, Massachusetts; and* [2]*Biophysics and Biomedical Modeling Division, US Army Research Institute of Environmental Medicine, Natick, Massachusetts*

Submitted 21 May 2020; accepted in final form 28 July 2020

Salgado RM, Caldwell AR, Coffman KE, Cheuvront SN, Kenefick RW. Endurance test selection optimized via sample size predictions. *J Appl Physiol* 129: 467–473, 2020. First published July 30, 2020; doi:10.1152/japplphysiol.00408.2020.—Selecting the most appropriate performance test is critical in detecting the effect of an intervention. In this investigation we *1*) used time-trial (TT) performance data to estimate sample size requirements for test selection and *2*) demonstrated the differences in statistical power between a repeated-measures ANOVA (RM-ANOVA) and analysis of covariance (ANCOVA) for detecting an effect in parallel group design. A retrospective analysis of six altitude studies was completed, totaling 105 volunteers. We quantified the test-retest reliability [i.e., intraclass correlation coefficient (ICC) and standard error of measurement (SEM)] and then calculated the standardized effect size for a 5–20% change in TT performance. With these outcomes, a power analysis was performed and required sample sizes were compared among performance tests. Relative to TT duration, the 11.2-km run had the lowest between-subject variance, and thus greatest statistical power (i.e., required smallest sample size) to detect a given percent change in performance. However, the 3.2-km run was the most reliable test (ICC: 0.89, SEM: 81 s) and thus better suited to detect the smallest absolute (i.e., seconds) change in performance. When TT durations were similar, a running modality (11.2-km run; ICC: 0.83, SEM: 422 s) was far more reliable than cycling (720-kJ cycle; ICC: 0.77, SEM: 480 s). In all scenarios, the ANCOVA provided greater statistical power than the RM-ANOVA. Our results suggest that running tests (3.2 km and 11.2 km) using ANCOVA analysis provide the greatest likelihood of detecting a significant change in performance response to an intervention, particularly in populations unaccustomed to cycling.

NEW & NOTEWORTHY This is the first investigation to utilize time-trial (TT) data from previous studies in simulations to estimate statistical power. We developed an easy-to-use decision aid detailing the required sample size needed to detect a given change in TT performance for the purpose of test selection. Furthermore, our detailed methods can be applied to any scenario in which there is an impact of a stressor and the desire to detect a treatment effect.

decision aid; exercise performance; hypoxia; test-retest reliability

INTRODUCTION

When assessing whether a particular experimental intervention/treatment (e.g., training practice, pharmacological agent, nutritional supplement) can alter endurance performance, typically quantified by a time-trial (TT) test, the choice of the most appropriate test (e.g., exercise mode and duration/length) is critically important. The sample size required to statistically detect a desired effect on TT performance is predicated on the size of the true effect relative to the variance (i.e., signal-to-noise ratio). There are numerous factors that influence variability in TT performance; often critical are environmental stressors, duration, and mode of exercise (2, 7, 12).

Many factors such as environmental stressors (e.g., high altitude or ambient temperatures) negatively impact the ability to complete a given endurance task compared with neutral environmental conditions (9, 10). The degree to which such conditions impair TT performance appears to vary (2, 6, 10, 17), and its magnitude increases in more severe conditions (i.e., higher elevations or temperatures) (2, 17). Therefore, the effect of an intervention that is used to mitigate decrements in endurance performance, such as a nutritional intervention to diminish the effects of altitude (4), may be difficult to detect when the outcome measure can vary greatly (i.e., low test-retest reliability) in different conditions. Although the reliability of various physical performance tests has previously been quantified (7, 12), its impact on a study's statistical power—the conditional probability of detecting a significant effect—has not been considered.

The use of the most appropriate statistical analysis is important when analyzing and interpreting any data. Studies that evaluate the efficacy of an intervention to attenuate the stressor-dependent decrements in performance commonly utilize a pre-post parallel group design (e.g., control vs. treatment group from preintervention to postintervention) (13, 14). Many will analyze these data with a between-within repeated-measures ANOVA (RM-ANOVA). However, because this analysis requires partial eta squared (η^2) or Cohen's *f* effect sizes, which are not intuitive compared with Cohen's *d* (for pre-post study designs), researchers may find that a power analysis for this statistical procedure can be challenging to perform. Furthermore, comparing change scores between a control and a treatment group with a RM-ANOVA can produce a biased result compared with an analysis of covariance (ANCOVA) (29), in which the baseline value is used as a covariate. With regard to study design, the use of ANCOVA may be more efficient and require a smaller sample size to detect an effect (e.g., increased statistical power) (30).

Recent developments within simulation software have made it considerably easier to perform power analyses for more complicated designs (20). For example, with a few modifications to the code in the "Superpower" R package (20), the power calculations can be performed for both RM-ANOVA and ANCOVA. However, to our knowledge, no study has

Correspondence: R. M. Salgado (roy.m.salgado.civ@mail.mil).

performed simulations of power using RM-ANOVA and ANCOVA utilizing common experimental designs to assess the efficacy of an intervention to improve an outcome measure, such as exercise performance.

To that end, we utilized existing TT performance data from different modes of exercise (cycling and running) performed at various altitudes. Such data provide us with a scenario in which an environmental stressor impairs performance, and we can then compare the required sample size needed to detect a given desired change in performance from an intervention for the purpose of test selection. Therefore, the purpose of this study was twofold: *1*) use TT variability (i.e., test-retest reliability) to calculate and compare the required sample size needed to detect a meaningful change in performance for each type of exercise test (i.e., duration and modality) at sea level and at altitude and *2*) compare the power of different statistical tests (RM-ANOVA and ANCOVA) for detecting a statistical effect. We hypothesize that the shorter-duration test (i.e., 3.2-km run vs. 11.2-km run and 720-kJ cycle) and the running modality (i.e., 3.2-km run and 11.2-km run vs. 720-kJ cycle) would have less variability and thus smaller sample size to detect a given theoretical change in endurance exercise performance. Using a common research design (pre-to-post parallel group design), we aimed to show that the ANCOVA would be more powerful than the RM-ANOVA for detecting a simulated treatment effect in any combination of duration and exercise mode at altitude. Finally, in summarizing our results, we plan to develop a user-friendly decision aid to assist researchers in a priori estimates of either sample size or exercise modality or both when making initial plans for a new study. Equally important, the methodological approach used in this investigation has practical application in almost any situation where there is an impact (small or large) of a stressor (e.g., load carriage, dehydration, high altitude, or ambient temperatures) and a desire to measure a potentially smaller effect of some type of intervention (pharmaceutical, nutritional supplement, training regimen).

METHODS

Overview of Studies

A retrospective data analysis was performed with TT performance data from eight studies contained within the US Army Research Institute of Environmental Medicine (USARIEM) Mountain Medicine Database. To be included in the analysis, TT performance must have been measured at least twice at sea level (for test-retest reliability) and once at altitude (2,500, 3,000, 3,500, 4,050, or 4,300 m; to measure the effect of altitude). If a study used an intervention, the treatment group was included in the analysis only if the performance values were not statistically different from the placebo (i.e., lack of an altitude $\times$ treatment interaction). If repeated measures were performed at altitude, only the first exposure was included; importantly, all exercise endurance performance tests included in the present analysis were completed within 72 h of exposure to altitude. Individuals who reported having had acute mountain sickness (AMS) [AMS-cerebral (AMS-C): $\geq$0.7 (28) or AMS-Lake Louise (AMS-LL): $\geq$3 (25)] were excluded because *1*) symptoms of AMS are thought to impair exercise performance and *2*) the time at which AMS was assessed before measuring exercise performance was not consistent among the studies. All volunteers were born at altitudes $<$1,500 m and resided at altitudes $<$1,200 m for at least 2 mo before participating in the study. Volunteers were briefed on the specifics of the study and provided verbal and written informed consent before participating. All studies were approved by the Institutional Review Boards at the USARIEM and/or the US Army Medical Research and Materiel Command. Investigators adhered to Department of Defense (DoD) Instruction 3216.02 and 32 CFR 219 on the use of human research.

Time-Trial Performance

Volunteers completed a running or cycling TT test at a self-selected pace. For the running modality, volunteers completed a 3.2-km or 11.2-km run TT test on either a track or a motorized treadmill with fixed grade (1% or 3%). During the treadmill TT, volunteers were free to increase or decrease the speed. For the cycling modality, volunteers completed a 720-kJ cycle test on an electronically braked cycle ergometer. The longer-duration 11.2-km run and 720-kJ cycle tests were chosen because the decrement in aerobic performance at high altitude would be greater compared with shorter-duration tests, whereas the 3.2-km run was chosen because this distance is the same as used in the US Army Fitness Test and thus ecologically valid. Volunteers were free to alter pedaling cadence and adjust workload. In both exercise modalities, volunteers were provided distance (running) or work completed (cycling) but were blinded to the speed or power output and elapsed time. The volunteers were instructed to complete the performance test as quickly as possible, and exercise tests were separated by at least 1 wk. As noted above, the TT tests were completed twice at sea level to account for practice effects.

Statistical Analyses

Effect of altitude. To describe the effect of altitude on TT performance, the mean change in time (s) was calculated from sea level to the target altitude and the standardized mean difference was expressed as Cohen's d_z:

$$\text{Cohen's } d_z = \frac{M_{\text{change}}}{\sqrt{\frac{\Sigma\left(I_{\text{change}} - M_{\text{change}}\right)^2}{N-1}}} \quad (1)$$

where M_{change} is the mean change in TT time, I_{change} is the individual change in TT time, and N is the total number of pairs (pre to post). The Cohen's d_z estimate was also corrected for bias with a bootstrap method (26).

Power analyses. A power analysis was performed for each exercise test to detect a theoretical change in TT performance of 5% and 20% for measures at sea level. First, the following reliability statistics were calculated: *1*) Pearson's correlation coefficient (r), *2*) intraclass correlation coefficient ($ICC_{3,k}$) (33), and *3*) the standard error of measurement (SEM) (33).

Next, with the reliability statistics, the estimate of the standard deviation of the change in scores was calculated as follows:

$$SD_{\text{change}} = \sqrt{2 \cdot SD_{TT}^2 - 2 \cdot r_{TT} \cdot SD_{TT}^2} \quad (2)$$

where SD_{change} is the standard deviation of the change score (in TT time), SD_{TT} is the standard deviation of the TT performance, and r_{TT} is the correlation between the repeated measurements of the TT performance at sea level.

Finally, the effect size, Cohen's δ_z, used for the power analysis (19), was determined by the following formula:

$$\text{Cohen's } \delta_z = \frac{M_{\text{TT Sea Level}} \cdot \%_{\text{change}}}{SD_{\text{change}}} \quad (3)$$

where $M_{\text{TT Sea Level}}$ was the average TT performance at sea level and $\%_{\text{change}}$ was the percent change in performance (i.e., 5% and 20%). The notation δ here is used because this standardized mean difference is referenced to a hypothesized population parameter, not an estimate (d) from a sample. Bias-corrected and accelerated bootstrap confidence intervals (95%) were calculated for each statistic with the "boot" package (5) with 5,000 replicates. All analyses assumed a

desired power of 80% ($\beta = 0.2$) with a significance level of $\alpha = 0.05$. These analyses were completed with the "pwr" package (5) within R (24).

Simulations for Power Analysis for Altitude Study

Using simulations, we determined the effect that different modalities (running vs. cycling) and statistical tests (RM-ANOVA vs. ANCOVA) would have on power in a study using a hypothetical treatment to attenuate altitude-induced decrements in aerobic performance. We simulated four separate scenarios for each exercise test assuming differing effects of altitude (an altitude-induced decrement in performance of 20% or ~3,400 m and 33% or ~3,800 m) and differing ergogenic benefit from the hypothetical intervention that theoretically attenuates the decrement by 5% and 20%. For the simulated study design we assumed a two-level between-subjects factor (e.g., Group: treatment and control groups) and a two-level within-subjects factor (e.g., Time: sea level and altitude), with TT performance as the dependent variable. The correlation between sea-level and altitude measurements was estimated from a meta-analysis via a linear mixed-effect model with the Hunter–Schmidt estimator (32) to account for attenuation in the correlation between repeated measurements due to altitude. As a note, very little heterogeneity, range $I^2 = 0–11.2\%$, was observed between studies with regard to the correlation between sea-level and altitude TT performance within each modality. The correlation between sea-level and altitude performance, from the meta-analysis, was 0.67 (0.49–0.86), 0.67 (0.47–0.87), and 0.86 (0.77–0.94) for the 720-kJ cycle, 11.2-km run, and 3.2-km run, respectively. In the RM-ANOVA, the Time × Group interaction was the statistical result of interest. For the ANCOVA analysis, the altitude TT performance was the dependent variable whereas the sea-level performance was a covariate with the effect of Group, two-level between-subjects factor (treatment and control group), as the statistical result of interest. To accomplish this task we utilized the "Superpower" R package (20). Furthermore, custom R functions were created to estimate power for the ANCOVA analysis. Because of the relatively complex nature of the analyses employed in this manuscript we have hosted the analysis scripts and synthetic data, which can be accessed at https://osf.io/gmz3a/ (see ENDNOTE).

RESULTS

Subject and Study Characteristics

Six studies from the USARIEM Mountain Medicine Database met the inclusion criteria, which resulted in 105 unacclimatized healthy sea level-native men ($n = 100$) and women ($n = 5$) included in this analysis (Table 1). All volunteers were unacclimatized healthy, fit US Army active-duty or college-aged individuals. Studies were completed in a hypobaric chamber [2 studies (1, 2); Natick, MA], in a hypoxia tent [1 study (3); Natick, MA], or at the Pikes Peak (PP) Research Laboratory [3 studies (4, 8, 23); Colorado Springs, CO, 4,300 m].

In four studies (1–3, 8) TT performance was assessed within ~3 h of arrival at the target altitude and was preceded by measurements of resting ventilation and acute mountain sickness and blood samples were collected. In the study by Andrew et al. (1), volunteers were given either *N*-acetylcysteine or placebo twice a day 2 days before and 2 days while at altitude. The study aim for Beidleman et al. (2) was to assess AMS and TT performance at various altitudes, and thus there was no study intervention. In a different study by Beidleman et al. (3), volunteers were divided into either intermittent hypoxic exposure ($P_{O_2} = 90$ mmHg) or "sham" hypoxia ($P_{O_2} = 148$ mmHg) for 3 h/day for 7 days. TT performance at altitude was completed before and after the intermittent hypoxic exposure. In the study by Fulco et al. (8), volunteers slept for 7.5 h/night for seven consecutive nights either in normobaric hypoxia (ambient O_2 progressively reduced from 16.2% to 14.4% from *night 1* to *night 7*) or in "sham" hypoxia (21.0% O_2) before altitude exposure. In one study (4), TT performance was conducted within 5 h of arrival at the 4,300 m altitude and was preceded by an 80-min steady-state exercise in which volunteers consumed either carbohydrate mixture or placebo. Muscle biopsy samples were collected before the steady-state exercise and after the TT test. In another study (23), volunteers received either autologous erythrocyte or placebo (saline) infusion before traveling to altitude and completed the 3.2-km run after 72 h of arrival at PP. Volunteers completed a peak oxygen consumption ($\dot{V}O_{2peak}$) and time-to-exhaustion exercise during the initial 2 days at altitude.

Effect of Altitude

Table 2 shows the effect of altitude on TT performance from the tests (3.2-km run, 11.2-km run, 720-kJ cycle), including the correlation (*r* for SL-ALT) for between performances at sea level and at various altitudes. On average, at altitudes ranging from 2,500 to 4,300 m, it took 33% longer to complete a given TT test (i.e., performance was impaired at altitude).

Reliability and Sample Size Estimations for Tests Performed at Sea Level

Table 3 shows the reliability statistics and sample size estimations for the TT tests at sea level. The relative SEM was smaller in both running modalities compared with the cycling modality. The reliability (ICC) of the running tests was also generally higher than the cycling test. Thus, the running tests required a smaller sample size to detect a given change in performance. The 11.2-km run resulted in the smallest sample size required to detect a given change in performance, which was likely driven by the smaller relative SEM. However, the

Table 1. *Volunteer characteristics separated by study*

Reference	Age, yr	Height, cm	Weight, kg	SL $\dot{V}O_{2peak}$, mL·kg^{-1}·min^{-1}	Sex, *n* (men, women)	Type of Time Trial
(1)	22 ± 3	176.1 ± 8.5	77.0 ± 12.0	51.1 ± 5.8	14, 1	11.2-km run
(2)	23 ± 4	176.8 ± 7.4	79.5 ± 12.4	43.9 ± 7.8	24, 2	720-kJ cycle
(3)	21 ± 3	176.7 ± 6.1	77.7 ± 12.2	48.3 ± 4.8	16, 0	720-kJ cycle
(4)	23 ± 6	176.5 ± 7.3	81.9 ± 13.9	51.6 ± 7.3	17, 0	3.2-km run
(8)	22 ± 4	172.1 ± 8.9	72.1 ± 9.5	46.4 ± 7.6	14, 2	11.2-km run
(23)	30 ± 3	177.7 ± 6.1	81.9 ± 6.0	53.6 ± 3.8	15, 0	3.2-km run

Values are expressed as means ± SD for $n = 105$ subjects. SL, sea level.

Table 2. *Time-trial performance tests from two modes of exercise performed at sea level and at various altitudes including sea level-altitude correlation and the standardized effect size of altitude on TT performance*

Exercise Mode	Altitude, m	*n*	Type of Time Trial	Time-Trial Performance, s				*r* for SL-ALT	Cohen's d_z†
				Sea Level	Altitude	Mean Change			
Cycling	2,500	14	720 kJ	4,957 ± 926	5,275 ± 1424	318 (35–879)	6.4%	0.88 (0.34 to 0.93)	0.39 (−0.18 to 0.76)
	3,000	7	720 kJ	4,969 ± 703	6,273 ± 1182	1,303 (776–1,949)	26.2%	0.71 (0.41 to 0.92)	1.29 (1.02 to 2.30)
	3,500	5	720 kJ	4,483 ± 650	5,402 ± 1712	919 (310–2,590)	20.5%	0.79 (0.68 to 0.90)	0.73 (0.63 to 0.98)
	4,300	16	720 kJ	4,622 ± 625	6,937 ± 1326	1,183 (848–1,697)	50.1%	0.45 (−0.09 to 0.76)	1.79 (1.40 to 2.68)
Running	3,500	15	11.2 km	5,020 ± 570	6,397 ± 756	1,376 (1,098–1,658)	27.4%	0.66 (0.21 to 0.86)	2.27 (1.78 to 3.12)
	4,050	16	11.2 km	4,626 ± 658	6,458 ± 921	1,833 (1,506–2,151)	39.6%	0.67 (0.26 to 0.83)	2.54 (2.11 to 3.58)
		15	3.2 km*	821 ± 77	1,161 ± 96	341 (319–363)	41.5%	0.89 (0.32 to 0.98)	6.98 (5.72 to 9.87)
		17	3.2 km	1,014 ± 148	1,516 ± 369	501 (410–616)	49.4%	0.76 (0.61–0.86)	1.74 (1.37 to 2.38)

Values are expressed as means ± SD; values in parentheses represent the 95% confidence intervals. ALT, altitude; SL, sea level. *Time-trial test was completed on a designated track. †Reported Cohen's *d* estimates were bias corrected

ICC between sea-level and altitude performance was highest in the 3.2-km run.

Simulations of Power Analysis for Altitude Study

The power analyses derived from the simulations are shown in Table 4**.** In general, to detect a hypothetical treatment effect of 5%, the ANCOVA approach to analyzing differences between treatment groups was more powerful (i.e., more likely to detect a treatment effect) than the RM-ANOVA. However, if the desire was to detect a treatment effect of 20%, the RM-ANOVA and ANCOVA provided similar statistical power. Overall, the running tests (3.2 km and 11.2 km) were more powerful, thus requiring a smaller sample size compared with the cycling test (720 kJ). Among the three tests, the 11.2-km run was the most powerful exercise test; this result was primarily driven by the higher correlation between sea level and altitude, which reduced the residual variance.

DISCUSSION

Selecting an accurate and reliable outcome measure, such as a TT performance test, is critical to the ability to detect the effect of an intervention. Thus, we aimed to *1*) use test-retest reliability to calculate and compare the required sample size needed to detect a meaningful change in performance for each TT test from sea level to altitude, in order to optimize test selection, and *2*) use simulations to compare the power of ANOVA to that of ANCOVA for detecting the effects of a potential treatment in attenuating the decrement in TT performance observed at altitude. Our main findings were that *1*) running modality was more reliable (Table 3) than the cycling modality and *2*) ANCOVA for pre-post designs offered greater statistical power. These findings have broad application. Although we utilized TT performance data performed at various altitudes, this methodological approach can be applied to a variety of stressful scenarios such as environmental heat stress, dehydration, or sleep deprivation in order to determine the appropriate performance test and the sample size required to assess the efficacy of an intervention (e.g., cooling vest, fluid replacement, caffeine) that is intended to attenuate the effect of a stressor.

A unique aspect of this study was that we used test-retest data (sea level and altitude) from each TT test to inform the power analyses. We also used simulations to compare power from common statistical analyses (RM-ANOVA and ANCOVA) used in parallel study designs. The methods used in the present study allowed us to compare, among the TT tests, the sample size needed to detect a theoretical effect (as a % change in TT performance). The information can be used to determine the most efficient, in terms of statistical power, outcome measure to track changes in performance among TT tests of

Table 3. *Reliability and sample size estimations for time-trial tests at sea level assuming $\alpha = 0.05$ and $\beta = 0.2$*

Modality	Type of Time Trial	ICC	SEM, s	Relative SEM, %	5% Increase		20% Increase	
					δ_z	Sample Size	δ_z	Sample Size
Cycling	720 kJ (n = 42)	0.77 (0.60–0.93)	480 (379–635)	10.1 (7.9–13.1)	0.18	237	0.73	17
Running	11.2 km (n = 31)	0.83 (0.62–0.94)	422 (293–540)	8.2 (6.1–10.9)	0.36	61	1.46	6
Running	3.2 km (n = 32)	0.89 (0.83–0.94)	81 (59–119)	8.6 (6.3–12.2)	0.20	190	0.81	13

Values in parentheses for the reliability statistics are bootstrapped 95% confidence intervals. ICC, intraclass correlation coefficient, reliability derived from repeated sea-level performances; relative SEM, standard error of measurement (SEM) relative to mean sea-level time-trial performance time; δ_z, effect size.

Table 4. *Required sample size per group for ANCOVA vs. ANOVA assuming $\alpha = 0.05$*

Type of TT Test	Effect of Altitude	Effect of Treatment*	ANCOVA Power		ANOVA Power	
			80%	95%	80%	95%
720-kJ cycle	20%	5%	243	402	258	429
		20%	17	27	17	29
	33%	5%	198	328	210	349
		20%	14	22	14	23
11.2-km run	20%	5%	71	120	74	123
		20%	6	8	6	9
	33%	5%	60	98	60	100
		20%	5	7	5	8
3.2-km run	20%	5%	154	254	247	409
		20%	11	18	17	27
	33%	5%	126	208	201	310
		20%	10	15	14	22

ANCOVA, analysis of covariance; TT, time trial. *Assumes an attenuation from placebo group, e.g., if there is a 20% increase from altitude alone, then a 5% treatment effect would result in a 1% attenuation in the increase from altitude (20% vs. 19% increase in control vs. treatment groups. respectively).

various durations and modalities. Previous studies have only reported the reliability statistics (7, 11) and thus are limited in practical application.

Contrary to our hypothesis, although the shorter-duration 3.2-km run was the most reliable test, it did not result in the smallest sample size to detect a treatment effect on endurance performance. Instead, our analysis indicates that the 11.2-km run offers the greatest statistical power for a given relative (percent) change in performance. The relative SEM of the 11.2-km run test was the lowest among the three TT tests (Table 3). Relative to average TT duration, the between-subjects variance was lower in the 11.2-km run, and therefore, for any given percentage change in performance, statistical power was higher. However, the lower between-subjects variance for the 11.2-km run can be the result of the type of subjects and similarities in the study design. Thus, these results may not be transferable to studies with volunteers of different fitness levels and experimental designs. We decided to express the change in performance from a treatment in relative units (i.e., 5% and 20%). We believe this approach allows for a fair comparison among different TT tests with the same units (e.g., time in seconds). This is mainly because absolute changes in performance (i.e., 90 s) would have a different impact on an 11.2-km run (relatively small change) compared with a 3.2-km run (relatively large change). Still, some caution is warranted before giving a broad recommendation to use the 11.2-km run over the 3.2-km run, particularly when considering the precision of a measurement. If the change in performance were instead expressed in absolute units (i.e., seconds or minutes), the 3.2-km run would clearly offer the greatest power. For example, if researchers are interested in detecting an absolute change in performance, such as 90 s, the 3.2-km run has the highest reliability and smallest between-subject variation and therefore would have the highest statistical power to detect a treatment effect.

Our findings show that when exercise test durations are similar a running modality (11.2-km run) is more reliable (sea-level ICC: 0.83 and SEM: 422 s) than a cycling modality (720-kJ cycle, sea-level ICC: 0.77 and SEM: 480 s; Table 3). These findings are in contrast to other analyses that reported no differences in reliability between the two modes of exercise (7, 12). It is plausible that the running tests were more reliable than the cycling test in this study because the volunteers included in the present analysis (primarily military personnel) were more familiar with running as opposed to cycling. Although the difference in exercise test duration at sea level between the 11.2-km run and the 720-kJ cycle is negligible (Table 3**;** 11.2-km run: 80 min vs. 720 kJ: 79 min), the larger ICC and smaller SEM of the 11.2-km run result in a smaller sample size. For instance, in this scenario the sample size required to detect a 20% change in performance decreases by 11 volunteers when choosing the 11.2-km run over the 720-kJ cycle test (Table 3).

Longer-duration TT tests may be affected to a greater degree (i.e., higher % decrement in performance) in more severe conditions such as at high compared with low altitude (9). Thus, it could be argued that a given desired treatment effect (e.g., 5%) to attenuate the impact of a stressor is more likely to be detected at, for example, higher altitudes, with a longer-compared with shorter-duration TT test. This may be the case, assuming the variance in a TT test is not affected by the stressor of interest and/or the effect of the stressor is so large that the increased variance has a limited effect on the standardized effect size (e.g., partial η^2). However, at least within our altitude data, exposure to such conditions introduces a new source of variance, as indicated by the lower correlation from sea level to altitude compared with repeated measures of performance at sea level, as well as the increased standard deviation at all TTs at altitude compared with sea level. Although we cannot say for certain, other situations such as the comparison of low-weight to heavy-weight load carriage or increased ambient temperatures may also introduce variance in TT data. Therefore, if the goal is to detect a treatment to attenuate the effect of a stressor, the more reliable test and not necessarily the one with a greater hypothetical effect is recommended.

An important outcome of this study was the development of an easy-to-use decision aid that can be used to inform researchers of the appropriate sample size for sea level and/or altitude studies in which the change in TT performance is the primary measurement (Table 3 and Table 4). For example, at sea level, using the 11.2-km run test, at least 6 volunteers are required to detect a 20% change in TT performance (Table 3). Similarly, assuming it takes 33% longer to complete an endurance task at altitude relative to sea level (i.e., impairment in performance) and a desire to detect a 5% attenuation (i.e., improvement) from a treatment, using an 11.2-km run requires a minimum of 60 volunteers per group ($N = 120$, power = 80%; Table 4). This is obviously a very large sample size and much larger than most studies in this research area. This illustrates that only relatively large effects of treatments (e.g., 20% attenuation) can be observed with the typical sample size obtained in altitude studies.

Our simulations also demonstrate that an ANCOVA is slightly more powerful than the RM-ANOVA. Van Breukelen (30) reported that the required sample size for the ANCOVA is only $\frac{(1+r_{pre-post})}{2}$ as large as that for a RM-ANOVA (22). This is an approximation, and this formula assumes homoge-

neity of variance. For the different tests in this study, this formula would estimate a 7.1%, 16.5%, and 16.2% smaller sample size required for the 3.2-km run, 11.2-km run, and 720-kJ cycling TT, respectively. In situations where the effects are large, e.g., a 33% effect of altitude with a 20% treatment effect, the benefit of the ANCOVA approach is diminished because the required sample size for the RM-ANOVA is already small. However, if the correlation between pre (sea level) and post (altitude) TT performance is lower than the estimates within our simulations, then the benefit of utilizing an ANCOVA is even greater. In any case, for a simple parallel group design, we recommend that researchers use the ANCOVA over the RM-ANOVA for the greater statistical power and a lower risk of bias (30, 31).

Application of Our Methodological Approach to Other Research Areas

This analysis included three endurance TT tests encompassing a wide range of durations (13.7 to 80 min) and the two most common exercise modalities (running and cycling) performed at sea level and various altitudes. Other laboratories may be interested in other stressors that degrade performance, populations with different fitness level or health status (i.e., clinical population), and/or different tests to assess performance than those included in our analysis. We also acknowledge that the test-retest reliability among laboratories will vary. Thus, with the test-retest reliability of the specific population of interest, using our calculations, simulations, and corresponding code (see ENDNOTE), we provide a framework for researchers to determine their own sample size needs and optimize their performance test selection for their given study design. Furthermore, the methods used in this present study are not limited to measures of physical performance, as they can also be applied to a variety of dependent variables (e.g., physiological, psychometric, blood biomarkers, etc.) across different disciplines of research.

Limitations

One limitation of this investigation is that the intraindividual variability in performance decrements (e.g., heterogeneous response to altitude) was not accounted for in the sample size estimations at altitude. Although there is some indication that the variance in TT performance becomes larger at higher altitudes, the RM-ANOVA is likely to be robust to this violation when a Greenhouse–Geisser adjustment is implemented (18). Another limitation of this study could be the type of performance task we chose to analyze as well as the stressor, e.g., altitude. Given that only fixed time or work criterion tests were examined, our findings may not be applicable to criterion tests without a fixed end, such as time to exhaustion. Indeed, data from Jeukendrup et al. (15) and others also support greater overall reliability in TT tests compared with time to exhaustion; thus the latter may not be similarly interpretable. The test-retest reliability of the TT tests in this study were determined from young, healthy, fit but primarily untrained individuals who completed the tests at sea level and various altitudes and therefore may not be extended to all populations and different stressors. Furthermore, because the volunteer characteristics (e.g., $\dot{V}O_{2peak}$, height) varied among the studies, these factors may influence the effect sizes and/or reliability statistics. Thus, researchers are suggested to use our sample size tables as general guidance and are cautioned not to directly apply them to their studies. Nevertheless, in many cases, researchers can use our methodological approach to complete their power analysis. In this report, we assume that researchers are interested in the statistical power of a study. If researchers are more interested in the precision of an effect size, which was out of the scope of this study, they can use "accuracy in parameter estimation" or AIPE (21). Rather than power, this approach plans for "assurance" that a confidence interval will be of a certain width by a given sample size per group. Nevertheless, AIPE can be applied to both RM-ANOVA and ANCOVA with the MBESS package in R (16).

Conclusions

This is the first investigation, to our knowledge, to use TT data from previous studies in simulations with the goal of estimating the sample size required to detect the theoretical effect of a treatment to attenuate the effect of a stressor on TT performance. As a result we developed an easy-to-use decision aid (Tables 3 and 4) detailing the required sample size needed to detect a change (5% and 20%) in TT performance for the purpose of performance test selection. Additionally, we demonstrated the power of various analysis procedures (RM-ANOVA and ANCOVA) to detect an effect within these specific study designs and conditions. From our simulations, we provide a detailed analysis procedure that can be applied to research using a similar study design but with different outcome measurements. Our findings demonstrate that TT tests using a running modality resulted in the most statistically powerful study design and that the ANCOVA was more advantageous, in terms of power, than the RM-ANOVA. However, because the volunteers included in this analysis are healthy, fit individuals who were more familiar with running as opposed to cycling, these findings may not be generalizable to other populations (e.g., untrained individuals or cyclists). Furthermore, when the desire is to detect a percent change in performance, the 11.2-km run is more powerful than the 3.2-km run. However, when trying to detect an absolute change, i.e., seconds or minutes, the 3.2-km run would be more likely to detect smaller changes in performance.

ACKNOWLEDGMENTS

The authors thank Nisha Charkoudian for careful review of the manuscript and Stefan M. Pasiakos for providing additional 3.2-km run TT performance data. Additionally, we thank Andrew D. Vigotsky and Matthew S. Tenan for statistical expertise and feedback on this manuscript.

DISCLAIMERS

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or reflecting the views of the Army or the Department of Defense. Any citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement or approval of the products or services of these organizations. Approved for public release; distribution is unlimited.

DISCLOSURES

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

AUTHOR CONTRIBUTIONS

R.M.S., S.N.C., and R.W.K. conceived and designed research; R.M.S. and A.R.C. performed experiments; R.M.S., A.R.C., S.N.C., and R.W.K. analyzed data; R.M.S., A.R.C., K.E.C., S.N.C., and R.W.K. interpreted results of

experiments; R.M.S. and A.R.C. prepared figures; R.M.S. drafted manuscript; R.M.S., A.R.C., K.E.C., S.N.C., and R.W.K. edited and revised manuscript; R.M.S., A.R.C., K.E.C., S.N.C., and R.W.K. approved final version of manuscript.

ENDNOTE

At the request of the authors, readers are herein alerted to the fact that additional materials related to this manuscript may be found at a website hosted by the authors, which at the time of publication they indicate is: https://osf.io/gmz3a/. These materials are not a part of this manuscript and have not undergone peer review by the American Physiological Society (APS). APS and the journal editors take no responsibility for these materials, for the website address, or for any links to or from it.

REFERENCES

1. **Andrew S, Grunbeck M, Muza SR, Beidleman BA, McClung JM, Lammi E, Staab JE, Fulco CS.** *N*-acetyl-cysteine does not improve time-trial performance during altitude exposure. New England Chapter of American College of Sports Medicine Annual Meeting. Providence, RI, November 3, 2011.
2. **Beidleman BA, Fulco CS, Buller MJ, Andrew SP, Staab JE, Muza SR.** Quantitative model of sustained physical task duration at varying altitudes. *Med Sci Sports Exerc* 48: 323–330, 2016. doi:10.1249/MSS.0000000000000768.
3. **Beidleman BA, Muza SR, Fulco CS, Jones JE, Lammi E, Staab JE, Cymerman A.** Intermittent hypoxic exposure does not improve endurance performance at altitude. *Med Sci Sports Exerc* 41: 1317–1325, 2009. doi:10.1249/MSS.0b013e3181954601.
4. **Bradbury KE, Berryman CE, Wilson MA, Luippold AJ, Kenefick RW, Young AJ, Pasiakos SM.** Effects of carbohydrate supplementation on aerobic exercise performance during acute high altitude exposure and after 22 days of acclimatization and energy deficit. *J Int Soc Sports Nutr* 17: 4, 2020. doi:10.1186/s12970-020-0335-2.
5. **Champely S, Ekstrom C, Dalgaard P, Gill J, Weibelzahl S, Anandkumar A, Ford C, Volcie R, Rosario H.** *Power Analysis Functions along the Lines of Cohen (1988)*. 2018. https://cran.r-project.org/web/packages/pwr/pwr.pdf.
6. **Chapman RF, Stager JM, Tanner DA, Stray-Gundersen J, Levine BD.** Impairment of 3000-m run time at altitude is influenced by arterial oxyhemoglobin saturation. *Med Sci Sports Exerc* 43: 1649–1656, 2011. doi:10.1249/MSS.0b013e318211bf45.
7. **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.
8. **Fulco CS, Muza SR, Beidleman BA, Demes R, Staab JE, Jones JE, Cymerman A.** Effect of repeated normobaric hypoxia exposures during sleep on acute mountain sickness, exercise performance, and sleep during exposure to terrestrial altitude. *Am J Physiol Regul Integr Comp Physiol* 300: R428–R436, 2011. doi:10.1152/ajpregu.00633.2010.
9. **Fulco CS, Rock PB, Cymerman A.** Maximal and submaximal exercise performance at altitude. *Aviat Space Environ Med* 69: 793–801, 1998.
10. **Galloway SD, Maughan RJ.** Effects of ambient temperature on the capacity to perform prolonged cycle exercise in man. *Med Sci Sports Exerc* 29: 1240–1249, 1997. doi:10.1097/00005768-199709000-00018.
11. **Hopkins WG.** Measures of reliability in sports medicine and science. *Sports Med* 30: 1–15, 2000. doi:10.2165/00007256-200030010-00001.
12. **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.
13. **Hursh DG, Baranauskas MN, Wiggins CC, Bielko S, Mickleborough TD, Chapman RF.** Inspiratory muscle training: improvement of exercise performance with acute hypoxic exposure. *Int J Sports Physiol Perform* 14: 1124–1131, 2019. doi:10.1123/ijspp.2018-0483.
14. **James CA, Richardson AJ, Watt PW, Willmott AG, Gibson OR, Maxwell NS.** Short-term heat acclimation improves the determinants of endurance performance and 5-km running performance in the heat. *Appl Physiol Nutr Metab* 42: 285–294, 2017. doi:10.1139/apnm-2016-0349.
15. **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.
16. **Kelley K.** *MBESS: The MBESS R Package. R Package Version 4.6.0*, 2019. https://cran.r-project.org/web/packages/MBESS/MBESS.pdf.
17. **Kenefick RW, Cheuvront SN, Palombo LJ, Ely BR, Sawka MN.** Skin temperature modifies the impact of hypohydration on aerobic performance. *J Appl Physiol (1985)* 109: 79–86, 2010. doi:10.1152/japplphysiol.00135.2010.
18. **Keselman HJ, Algina J, Kowalchuk RK.** The analysis of repeated measures designs: a review. *Br J Math Stat Psychol* 54: 1–20, 2001. doi:10.1348/000711001159357.
19. **Lakens D.** Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol* 4: 863, 2013. doi:10.3389/fpsyg.2013.00863.
20. **Lakens D, Caldwell A.** *Simulation-Based Power-Analysis for Factorial ANOVA Designs*, 2019. https://osf.io/pn8mc/.
21. **Maxwell SE, Kelley K, Rausch JR.** Sample size planning for statistical power and accuracy in parameter estimation. *Annu Rev Psychol* 59: 537–563, 2008. doi:10.1146/annurev.psych.59.103006.093735.
22. **Overall JE, Doyle SR.** Estimating sample sizes for repeated measurement designs. *Control Clin Trials* 15: 100–123, 1994. doi:10.1016/0197-2456(94)90015-9.
23. **Pandolf KB, Young AJ, Sawka MN, Kenney JL, Sharp MW, Cote RR, Freund BJ, Valeri CR.** Does erythrocyte infusion improve 3.2-km run performance at high altitude? *Eur J Appl Physiol Occup Physiol* 79: 1–6, 1998. doi:10.1007/s004210050465.
24. **R Core Team.** *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2019. https://www.r-project.org.
25. **Roach RC, Hackett PH, Oelz O, Bärtsch P, Luks AM, MacInnis MJ, Baillie JK, Achatz E, Albert E, Andrews JS, Anholm JD, Ashraf MZ, Auerbach P, Basnyat B, Beidleman BA, Berendsen RR, Berger MM, Bloch KE, Brugger H, Cogo A, Costa RG, Cumpstey A, Cymerman A, Debevec T, Duncan C, Dubowitz D, Fago A, Furian M, Gaidica M, Ganguli P, Grocott MP, Hammer D, Hall D, Hillebrandt D, Hilty MP, Himashree G, Honigman B, Gilbert-Kawai N, Kayser B, Keyes L, Koehle M, Kohli S, Kuenzel A, Levine BD, Lichtblau M, Macdonald J, Maeder MB, Maggiorini M, Martin D, Masuyama S, McCall J, McIntosh S, Millet G, Moraga F, Mounsey C, Muza SR, Oliver S, Pasha Q, Paterson R, Phillips L, Pichon A, Pickerodt PA, Pun M, Rain M, Rennie D, Ri-Li G, Roy S, Verges S, dos Santos TBC, Schoene RB, Schoch OD, Singh S, Sooronbaev T, Steinback CD, Stembridge M, Stewart G, Stobdan T, Strapazzon G, Subudhi AW, Swenson E, Roger Thompson AA, van Patot MT, Twomey R, Ulrich S, Voituron N, Wagner DR, Wang S, West JB, Wilkes M, Willmann G, Yaron M, Zafren K; Lake Louise AMS Score Consensus Committee.** The 2018 Lake Louise Acute Mountain Sickness Score. *High Alt Med Biol* 19: 4–6, 2018. doi:10.1089/ham.2017.0164.
26. **Rousselet GA, Wilcox RR.** Reaction times and other skewed distributions: problems with the mean and the median. *Meta-Psychology* 4: MP.2019.16, 2020. doi:10.15626/MP.2019.1630.
28. **Sampson JB, Cymerman A, Burse RL, Maher JT, Rock PB.** Procedures for the measurement of acute mountain sickness. *Aviat Space Environ Med* 54: 1063–1073, 1983.
29. **Senn S.** Change from baseline and analysis of covariance revisited. *Stat Med* 25: 4334–4344, 2006. doi:10.1002/sim.2682.
30. **Van Breukelen GJ.** ANCOVA versus change from baseline: more power in randomized studies and more bias in nonrandomized studies. *J Clin Epidemiol* 59: 920–925, 2006. doi:10.1016/j.jclinepi.2006.02.007.
31. **Vickers AJ, Altman DG.** Statistics notes: analysing controlled trials with baseline and follow up measurements. *BMJ* 323: 1123–1124, 2001. doi:10.1136/bmj.323.7321.1123.
32. **Viechtbauer W.** Conducting meta-analyses in R with the metafor package. *J Stat Softw* 36: 1–48, 2010. doi:10.18637/jss.v036.i03.
33. **Weir JP.** Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 19: 231–240, 2005. doi:10.1519/15184.1.