Basic statistical considerations for physiology: The journal Temperature toolbox

Article · June 2019

DOI: 10.1080/23328940.2019.1624131

2 authors, including:

Aaron Caldwell
U.S. Army Research Institute of Environmental Medicine
50 PUBLICATIONS **103** CITATIONS

SEE PROFILE

Temperature
MULTIDISCIPLINARY BIOMEDICAL JOURNAL
Volume 2 - Issue 3 - July/August/September 2015

Special Issue: Japanese Thermal Physiology
Guest Editors: Osamu Shido, MD, PhD

Taylor & Francis

Temperature

ISSN: 2332-8940 (Print) 2332-8959 (Online) Journal homepage: https://www.tandfonline.com/loi/ktmp20

Basic statistical considerations for physiology: The journal *temperature* toolbox

Aaron R. Caldwell & Samuel N. Cheuvront

To cite this article: Aaron R. Caldwell & Samuel N. Cheuvront (2019): Basic statistical considerations for physiology: The journal *temperature* toolbox, Temperature, DOI: 10.1080/23328940.2019.1624131

To link to this article: https://doi.org/10.1080/23328940.2019.1624131

TEMPERATURE
https://doi.org/10.1080/23328940.2019.1624131

Taylor & Francis
Taylor & Francis Group

COMPREHENSIVE REVIEW

Basic statistical considerations for physiology: The journal *temperature* toolbox

Aaron R. Caldwell [a] and Samuel N. Cheuvront[b]

[a]Exercise Science Research Center, University of Arkansas–Fayetteville, Fayetteville, NC, USA; [b]Biophysics and Biomedical Modelling Division, US Army Research Institute of Environmental Medicine, Natick, MA, USA

ABSTRACT

The average environmental and occupational physiologist may find statistics are difficult to interpret and use since their formal training in statistics is limited. Unfortunately, poor statistical practices can generate erroneous or at least misleading results and distorts the evidence in the scientific literature. These problems are exacerbated when statistics are used as thoughtless ritual that is performed after the data are collected. The situation is worsened when statistics are then treated as strict judgements about the data (i.e., significant versus non-significant) without a thought given to how these statistics were calculated or their practical meaning. We propose that researchers should consider statistics at every step of the research process whether that be the designing of experiments, collecting data, analysing the data or disseminating the results. When statistics are considered as an integral part of the research process, from start to finish, several problematic practices can be mitigated. Further, proper practices in disseminating the results of a study can greatly improve the quality of the literature. Within this review, we have included a number of reminders and statistical questions researchers should answer throughout the scientific process. Rather than treat statistics as a strict rule following procedure we hope that readers will use this review to stimulate a discussion around their current practices and attempt to improve them. The code to reproduce all analyses and figures within the manuscript can be found at https://doi.org/10.17605/OSF.IO/BQGDH.

ARTICLE HISTORY
Received 16 November 2018
Revised 19 May 2019
Accepted 21 May 2019

KEYWORDS
Statistics; metascience; NHST; power analysis; experimental design; effect sizes; open science; nonparametric; preregistration; bootstrapping; optional stopping

"A statistician is a part of a winding and twisting network connecting mathematics, scientific philosophy, and other intellectual sources-including experimental sampling- to what is done in analyzing data, and sometimes in gathering it – John Tukey [1, p.225]".

Introduction

Statistics provides tools that allow researchers to make sense of the vast amount of information and data they collect. However, statistics can be a daunting challenge for many scientists. Few physiologists receive a formal education in statistics and for many formal mathematics or statistics training ended during their undergraduate studies [2]. However, statistics remain a key part of the research process because they allow researchers to infer their results to broader populations. Therefore, it is imperative for researchers to have a basic understanding of the underlying principles behind the statistical techniques they commonly use so they can make informed and accurate inferences in their research.

Statistics can *help* form conclusions but cannot replace good scientific reasoning and thought. Even among scientists who receive extensive statistical training, a number of biases and gross misunderstanding about statistics persist [3]. It appears the confusion arises when statistics are used as a ritual that provides the answers after the data are collected. Instead, researchers should utilize statistical principles to guide their thinking and reasoning before, during, and after data collection. Poor statistical practices have little to do with selecting the "right" statistical procedure but far more to do with a lack of thinking on the part of the researcher [3]. Data analysis is now rapid, easy and flexible due to modern statistical software [4], but this can contribute to ritual use of statistics where the numbers go in one end and a simple "yes" or "no" comes out the other.

Researchers can easily be lead to false or erroneous discoveries by the tendency to see patterns in random data, confirmation bias (only accepting information in favour of the hypothesis), and hindsight bias (explaining the data as predictable after it

CONTACT Aaron R. Caldwell arcaldwell49@gmail.com

is collected) [5]. Every statistical analysis gives researchers a number of potential decisions they can make (e.g., collect more observations, data transformations, exclusion of outliers, etc) and all of these decisions collectively give researcher's "degrees of freedom" in the data analysis [6]. This can make it incredibly easy to find a "significant" result when there is nothing but statistical noise [7,8]. For all of the above reasons, it is important to have a statistical analysis plan in place, *prior to data collection* to limit the researcher degrees of freedom (the flexibility to change analysis plan) and create more credible research practices [6,9].

This review is intended to be a guide for physiologists to avoid common pitfalls and mistakes as they plan, conduct, and analyse their studies. One should not use or cite this review as a simple justification for any one procedure or approach over another. Instead, we have posed questions throughout this review and if a study is appropriately designed and analysed then researchers should be able to adequately provide answers to these questions. We encourage readers to utilize this review as a brief, introductory guide to help one think about statistics to ensure their study is executed and analysed in the best way possible.

A quick note on null hypothesis significance testing and p-values

Currently, null hypothesis significance testing (NHST) is the predominate approach to inference in most scientific fields. In particular, environmental and occupational physiologists, whether they realize it or not, rely upon NHST which in large part is based on Jerzy Neyman and Egon Pearson's framework for inference [10–12]. In this paradigm, the data are collected and then the scientist must decide between two competing hypotheses: the null and the alternative. In essence, we collect a sample (a group of participants) from a population (the group that the researcher is trying to study), assuming we are interested in detecting a relationship or difference of *at least* a certain magnitude. After the data are collected, researchers use statistical test(s) to see if the observed difference or relationship is common, assuming the null hypothesis is true. In many cases, the null hypothesis is a statement that no difference or relationship exists (i.e., nil hypothesis). However, the null hypothesis can take the form of a variety of statements. For example, a null hypothesis could be that cold-water immersion does *not* cool a heat-stroke patient at least 0.05 °C/min faster than ice-sheet cooling (i.e., a minimum effect hypothesis).

The NHST utilizes *p*-values to help make decisions about the null and alternative hypotheses. The *p*-value provides evidence as to how uncommon this relationship or difference would be assuming the null hypothesis is true with data *at least as extreme* as was observed. While *p*-values below the designated error rate (typically 5%) are called "significant" it does not mean that something worthwhile has been discovered or observed. Significance in this context means that the statistical computation has *signified* something is peculiar about the data and requires greater investigation [13, p.98–99]. *P*-values *do not* provide direct evidence for or against any alternative hypothesis, and do not show the likelihood that an experiment will replicate [3]. Contrary to popular belief [3], all *p*-values are equally likely if there is no relationship or difference (Figure 1(a)), and higher *p*-values *do not* indicate stronger evidence for the null hypothesis. Instead, *p*-values are meant to prevent researchers from being deceived by random occurrences in the data by rarely providing significant results if the null hypothesis is true [14].

There are two types of errors to consider when utilizing NHST. The error rate for declaring an effect exists when there is no effect, a false positive, is the cut-off for significance (alpha level). This is also called the type I error rate. Most researchers utilize an arbitrary 0.05 alpha level (5%), but lower or higher alpha levels can be appropriate given the proper justification [15]. Neyman & Pearson also advocated for considering type II error, or the erroneous conclusion that the null hypothesis is true (false negative; Figure 1(b,c)). The distribution of *p*-values (Figure 1) varies depending on power, or the ability to detect an effect of a certain magnitude. The smaller the effect a researchers is trying to detect, the greater the sample size will need to be to have adequate power (discussed in greater depth in part 1). More so, sufficient power to detect the effect size of interest must be achieved in order to make strong conclusions about the hypothesis of interest. This approach has many of the desirable qualities of philosopher Karl Popper's falsification approach to scientific theories [16,17]. Essentially,

Table 1. Common statistical tests.

Tests	Description	Assumptions
Chi-square	• Applied to categorical data; test of frequencies	• Categorical or frequency data • Observations are independent • 2 or more categories or groups
t-Tests	• Used to evaluate differences • One sample, Dependent (within subjects) or Independent (between subjects) samples tests	• Parametric: data are normally distributed • Homogeneity of variance • No leverage points
Two one-sided tests (TOST)	• Essentially two t-tests • Tests for equivalence between groups, time points, or for correlations	• Same as regular t-tests • Equivalence bounds should be decided upon a priori [51]
Multiple Regression	• Linear (continuous) and logistic (dichotomous) • Extension of simple correlation except a number of variables (predictors) can be used to predict another variable (outcome/criterion)	• Parametric: residuals are normally distributed • Near constant variance (homoscedasticity) • Limited multi-collinearity between predictors
Analysis of Variance (ANOVA)	• An extension of linear regression • Test for mean differences among multiple (>2) groups ○ Post-hoc/pairwise comparisons tests should be decided upon a priori • Acceptable post-hoc or pairwise comparisons include: Bonferroni, Holm-Bonferroni, false discovery rate¥, Games-Howell*, Duncan*, Dunnett*, Tukey*, and Scheffe*	• Data are continuous • Independent variables (i.e., groups) are categorical not continuous • Data are random sampled and residuals are normally distributed • No leverage points
Hierarchical or Mixed Linear Models	• Advanced statistical designs typically used when data is nested (hierarchical) or when there are random and fixed factors (mixed) • Requires extensive statistical training ○ Physiologists should consult a statistician	• Linear relationship between variables • Residuals are normally distributed • Near constant variance (homoscedasticity)

Nonparametric and Robust tests	• Utilized when data violate assumptions required for parameter estimation
	  ○ Typically used when data do not meet assumption of normality
	  ○ Tests include Mann-Whitney (2 independent samples), Wilcoxon signed rank (paired samples), tests, Kruskal-Wallis (one-way; >2 groups), Friedman (one-way; repeated measures), M-estimators, and permutation/randomization tests (variety of designs)
	  ○ Post-hoc and pairwise comparisons may include a variety of procedures utilizing trimmed means or bootstrapping procedures [113,p.316–331]
		• Each tests has a specific set of assumptions
		• Data transformations may be necessary [61]

¥ ¯ Does not control familywise error rate
* – Only valid for between subjects (i.e., not repeated measures) designs

you establish a theory, collect data to strongly test that theory, and come to a conclusion based on your data. According to this framework, deciding the appropriate power (type II error rate) and significance level (type I error rate) is up to the researcher and the costs associated with committing each type of error [15].

More in depth reviews of the guiding principles of hypothesis significance testing exist [17–19], and this approach is not without its flaws and criticisms [20,21]. In many cases a Bayesian [22–24] or Likelihood [25,26] approach to hypothesis tests may be more appropriate depending on the research question and the information the researcher is wanting to obtain from the study at hand [27]. Furthermore, hypothesis tests could be abandoned entirely in favour of an estimation approach which focuses on effect sizes and

Table 2. Questions and checklist for designing experiments.

Questions to ask yourself
● What are the underlying theories for your research? ○ What do these theories predict? ○ What would falsify these theories? ● What do you hypothesize will happen in your experiment? ○ Do you have directional hypotheses? ○ Are you predicting no differences (i.e., equivalence or non-inferiority)? ● Do you have accurate measurements? ● What are possible confounding variables in your experiment? ● What should be your power and significance level? ○ What is the smallest effect size of interest?
Checklist
● Review the literature; establish theories and hypotheses to test these theories ● Ensure your tools and techniques have a reasonable measurement error ● Decide on a statistical analysis plan (be as specific as possible) ○ Determine the acceptable alpha (significance) and beta (power) error rates ○ Determine sample size (power analysis) ○ Include contingency plans for potential problems with the data (e.g., violation of assumptions of normality) ● Preregister your hypotheses, data collection methods, and statistical analysis plans. This can be done at osf.io, or aspredicted.com

confidence intervals [28]. We strongly recommend that readers take the time to read more about NHST or any approach to statistical inference they plan to utilize for their research.

Table 3. Questions and checklist for data collection.

Questions to ask yourself
● Is the data being collected accurately? ○ Are there new, unanticipated sources for error? ● Are you interested in parameter estimation or statistical significance? ○ Are you more interested in accuracy of the parameter or significance? ● What is an adequate sample size? ○ Do you want to have interim analyses?
Checklist
● Determine criteria for stopping the experiment(s) ○ Decide upon the parameter estimation or sequential analysis approach

Table 4. Questions and checklist for data analysis.

Questions to ask yourself
● What is the a priori data analysis plan? ○ What details were included in the study's preregistration? ● If someone was given your dataset would they understand what the variables are and how ● Are the data entered correctly, or are there potential errors? ● Does the data violate the assumptions for my statistical tests? ○ Are there outliers? Do they affect the outcomes of the analysis ○ Can you assume normality? ○ Are robust or non-parametric tests preferred? ● How will effect sizes be calculated?
Checklist
● Create a detailed data entry and maintenance procedure ○ Create a codebook that details the variable names ● Double check the data to ensure there are no erroneous data points ● Explore the descriptive statistics separated by group and data point ● Determine if the data meets the assumptions for the chosen inferential statistics ● Perform statistical analysis most appropriate for your data, your hypotheses, and fits with your preregistration ● Calculate effect size estimates ○ Ensure the estimate is appropriate and assumptions are met ● When appropriate, calculate confidence intervals around effect size estimates

Moving forward

The remainder of this review we have separated into four parts to reflect the four major portions of most research endeavours: designing the experiment, data collection, data analysis, and disseminating the results. All four parts are equally important for appropriate statistical practices. The purpose of this review is to encourage readers to utilize statistical thinking at every stage of a study, and apply these concepts to their own research.

Part 1: Experimental design

The designing of a study is arguably the most important part of the research process. Without a high-quality experimental design all other concerns are moot. If the data are inaccurate or the design is problematic, then the statistical analysis will *always* produce misleading results. Bad experimental designs no matter how they are analysed, will produce statistics of little worth. To put it another way, if you put flawed or garbage data

Table 5. Questions and checklist for the dissemination of results.

Questions to ask yourself
● How can I make these results easy to reproduce and replicate?
● How should I share the data or make the data availability to others?
○ What legal or ethical obligations do I have for making the data available?
● If I am sharing the data, does the codebook contain enough detail for someone else to understand my dataset?
● In order for my peers to verify my results, do I need to post my analysis scripts (e.g., code)?
○ Is my code annotated in enough detail for someone unfamiliar with my dataset and analysis plan to understand what was performed and why?
● Are the statistical results described in way that gives an appropriate level of uncertainty?
○ Are the limitations and assumptions of the techniques adequately discussed?
● Is the data visualized (figures) constructed in way that is appropriate?
Checklist
● Make data and code available (when appropriate) through services such as Open Science Framework (https://osf.io) or through Figshare (https://figshare.com)
● Provide information about the study design in great enough detail that experiments can be replicated and statistical analyses can be reproduced
● Create data visualizations that include individual data points and appropriate summary statistics
○ With larger samples, provide some type of visualization of the distribution
● Provide detailed information on the type(s) of statistical analysis utilized and if the assumptions of these tests were stratified
○ If the assumptions are not met, discuss how these may possibly affect the results and conclusions
● Discuss the uncertainty of your results, what limitations the study may have, and how future studies can investigate these theories and hypotheses further
○ Provide confidence intervals around effect size estimates

into a statistical analysis, then the analysis will only produce garbage information. Properly designing an experiment is essential, and this process should ensure the validity, reliability, and replicability of a study.

Develop strong theories and test them with strong hypotheses

In order to have good statistical hypotheses, strong theories need to be developed. Therefore, the start of any research endeavour should begin with a review of the relevant scientific literature. Researchers can then develop their own theories, hypotheses, and predictions regarding the physiological phenomena of interest. Theories can even be generated from simple observations and then can be formalized with data collection [29]. After a study or series of experiments, the evidence for or against any formal theory can be gleaned from the statistical analyses performed.

The process of good theory and hypothesis building is not easy. Theories or hypotheses should have a high degree of *verisimilitude*, or rather, should explain the data and previous observations while simultaneously making specific predictions about the future [30]. In addition, research should produce theories that are *prohibitive* or at least *falsifiable* [30–32]. There should be criteria that debunks or invalidates the theory. If a theory does not have any way to be disproven, or researchers only seek to confirm the theory, then the theory is pseudo-scientific [30].

Environmental and occupational physiologists should also consider the multi-factorial nature of physiological responses to environmental extremes when building theories [33]. In extreme environments individuals are exposed to a many stressors (e.g., at high altitude exposure to cold and hypoxia), and these stressors may interact with one another. This interaction may have an additive, synergistic, or even antagonist net effect on the outcomes of interest [33]. When possible, researchers should aim to have theories that explain the multi-factorial responses to the environment rather than responses to isolated phenomena.

Researchers should aim to design studies that are strong, or "severe", tests of their theories and hypotheses [34]. When studies are not severe tests then the statistical analyses are not really providing any worthwhile information about said theory. If a theory cannot be falsified, or at least find flaws in the theory, then there is only bad evidence for that theory because no test of the theory has occurred [16,p.5]. If the theory "survives" a severe test then the researcher can consider this as corroborative evidence. Designing studies to be severe tests of theories is difficult because researchers run the risk of finding conflicts in widely held beliefs. Studies are often weak and safe tests of a theory, and the weak or safe predictions made by researchers are often the object of the most scathing criticisms of

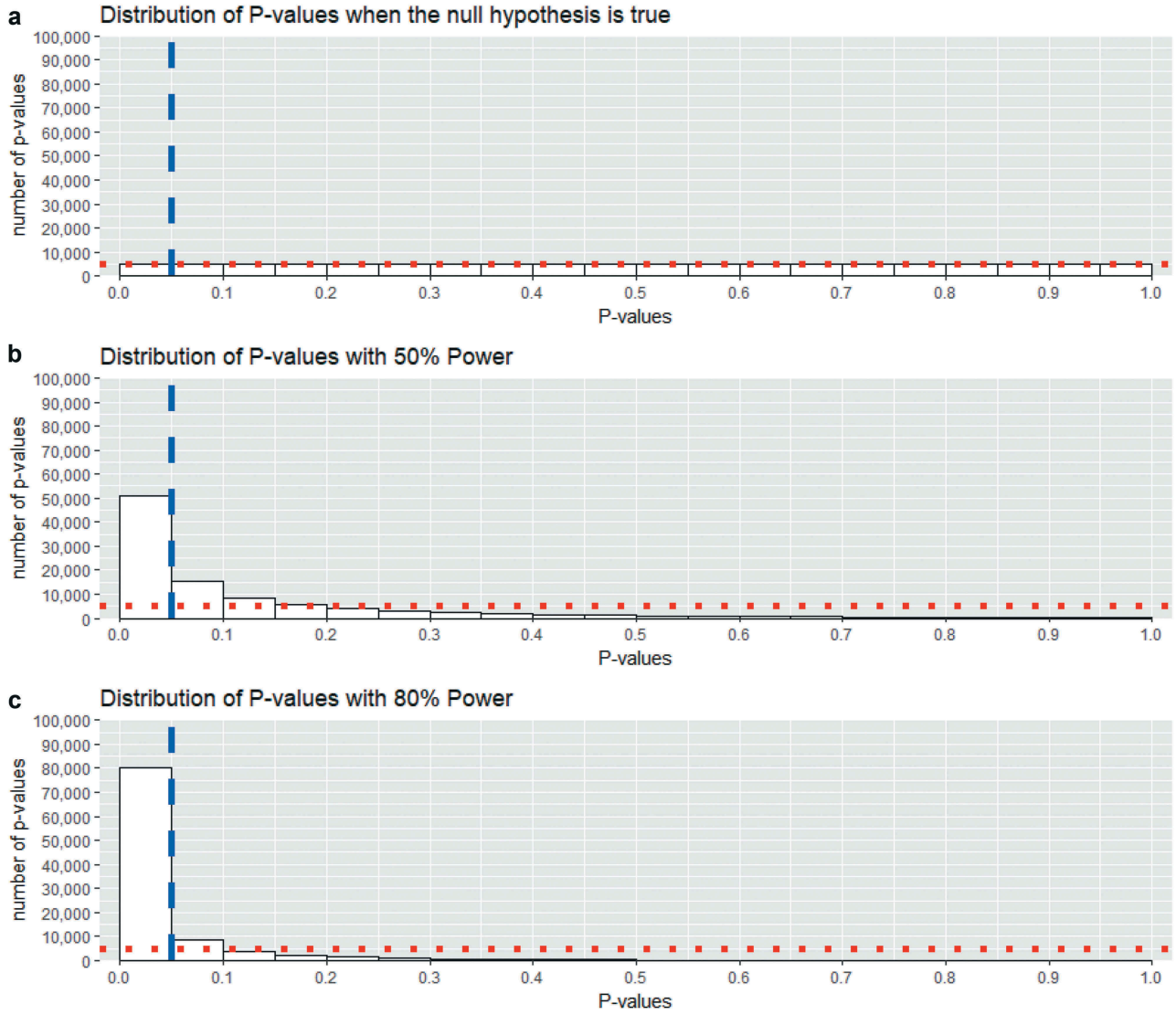


Figure 1. Results of a simulation (100,000 repetitions) demonstrating the distribution of p-values under the under null hypothesis (a), and when the null hypothesis is false when there is 50% power (b), and 80% power (c). The dotted horizontal red line indicates 5% of p-values in the simulation (5000 occurrences) per bin (5%) of the histogram. The dashed vertical blue line indicates the significance cut off 0.05. The 5000 p-values less than 0.05 in panel A are a type I error while the p-values greater than 0.05 in panels b & c are a type II error.

modern research [32]. In order to have severe tests, researchers should refine their theories, and hypotheses, to have *specific* falsifiable predictions about a physiological phenomenon. Severe tests of theories involve specific predictions about the direction and magnitude of such effects.

In environmental physiology, researchers can form strong theories and develop strong tests of these theories. For example, aerobic fitness has historically been considered a strong modulator of thermoregulation [35]. Consequently, relative exercise intensity has been used to normalize thermoregulatory comparisons between groups of high and low fitness. However, simple but clever challenges to this widely practiced between-group norming method, using a multi-factorial model, now indicate this modulatory effect likely does not exists or is smaller than originally predicted [36]. The study by Jay et al. [36], can be viewed as evidence of a "degenerative research programme" [37] or, to put it another way, the theory now explains less about thermoregulation than originally predicted. Further, novel methods for comparing time-dependent thermoregulatory responses for unmatched groups [38] can now be used to test other traditional assumptions about the roles

that body composition, sex, or age play in modulating thermoregulation [39]. At first, this aerobic fitness theory had a strong face validity, but strong tests of this theory [36,38] demonstrate that many predictions made by the initial theory do not hold up under scrutiny.

Measurement error

Measurement error is the difference between a measured quantity and the actual quantity. It is comprised of random errors and systematic errors. It is useful to understand the difference among error types so that they can be controlled (when possible) and their effect on the data quantified. A number of different statistics can quantify measurement error, but each are appropriate for certain situations.

The coefficient of variation (CV; one standard deviation divided by the mean) is one commonly used metric [40] as it includes the two most common statistics used to summarize probability distributions. The standard deviation is usually expressed as a percentage of the mean (%CV) and is key to the study of biological variation and useful for estimating sample size, effect size, and statistical probabilities [41].

A simple weighing scale is one of the most common pieces of equipment used in medicine and science, it is an essential piece of equipment for the study of sweating and hydration [41,42], and it affords a practical way of understanding measurement error in conjunction with the %CV. Random scale error may be intrinsic or extrinsic. For example, intrinsic random scale error occurs when the operating limits on a digital scale exhibit fluctuations in the last significant digit even when an inanimate object is weighed repeatedly; this worsens when a living specimen is weighed. Extrinsic random scale errors occur when there is lack of control over human factors such as ingestion or excretion of fluids, clothing, or items left inside pants pockets.

A systematic scale error, on the other hand, can be constant or proportional. Scales that are out of calibration are common in medical offices and may read ± 2 kg beyond the actual value [43]. This is an example of constant error. If for some reason the error increased or decreased when weighing larger or smaller people, the error would be proportional. Intrinsic random scale errors are unavoidable but measurable. Extrinsic random scale errors should be minimized to the greatest degree possible. Systematic errors can and should be avoided entirely or at minimum recognized and corrected, as is recommended for ingestible temperature pills [44]. An important caveat to thermal biologists is that the %CV should not be calculated for variables with an arbitrary zero origin, such as temperature. The variation in body temperatures measured in ^{0}C and ^{0}F will have the same variation but the %CV will change as a function of the arbitrary zero point [40].

As an example, the %CV for body weight measured day-to-day is ~1% (~0.70 kg) [41] when eliminating systematic errors and minimizing extrinsic random scale errors. If daily drinking adequacy was assessed using first morning body weight losses as a surrogate for water loss, a difference ≥ 1.60 kg would be required to reach statistical significance at the 95% confidence level (assuming unidirectional Z-score) [41]. On the other hand, an acute loss of body weight from sweating, with systematic and extrinsic random scale errors all but eliminated [42], reduces the %CV to intrinsic random scale error alone (often 0.05 kg) and a difference of just under 0.12 kg reaches statistical significance using the same probability assumptions. Both differences may be considered the smallest effects worth detecting when studying chronic or acute changes in body water, respectively, and illustrate one way in which the %CV can play an important role in designing experiments.

Selecting the appropriate inferential statistics

It is important to remember that each experimental design will have specific statistical tests that are appropriate for analysing the type of data that is collected. The first step in this process is deciding upon which outcomes are primary or secondary measures of interest. This is important because your primary outcome will determine the planning of the study (i.e., power analysis) and protects against cherry picking what results constitute evidence for your hypothesis. Moreover, each statistical test has a number of assumptions that need to be satisfied in order for those tests to be valid or informative. Researchers also have a number of analytic options. When left unchecked, researchers can run numerous analyses of the same outcome

measure, but are led astray by more attractive or positive results while ignoring the negative or null results [6–8,45,46]. Further, in studies with multiple outcomes measures (e.g., rectal temperature, thermal strain, inflammatory markers, etc) it can be easy to find at least a few significant results among all the outcome measures. Therefore, researchers should define the primary and secondary outcomes then determine what statistical test, or tests, are appropriate for a particular outcome measure *prior* to data collection [47].

First, the statistical analysis can change based on the type (differences vs. equivalence) and direction (greater vs. less) of the hypotheses. In most studies, researchers hypothesize that there is a difference, and then use statistical techniques, such as analysis of variance (ANOVA) or t-test, to test this prediction against the null, or rather nil, hypothesis that the true difference is zero [21,48]. In order to provide evidence that there are no differences, or at least that the differences are so small they are not relevant, researchers should use equivalence testing [49–53]. For simple designs (i.e., two group comparisons), equivalence testing can be accomplished with two one-sided tests (TOST) [54]. The TOST procedure simply consists of two one-sided t-tests against an upper bound and lower bound. For example, a research may hypothesize that sweat rates are approximately equivalent between midfielders and forwards on a soccer/football team, and sets the equivalence bounds to ± 0.25 L/h. Therefore, the TOST would test that the difference is more than –0.25 L/h, but less than 0.25 L/h. If the two t-tests in the TOST procedure are significant ($p < .05$), then the researcher can reasonable conclude that the difference is statistically smaller than any worthwhile effect, or, in other words, practically equivalent.

For more advanced designs, such as those with multiple groups or repeated measures there are textbooks [53,p.219–231] dedicated to the topic and free statistical packages to aid with the calculation of these statistics [54,55]. In these cases, traditional hypothesis tests are not appropriate because a non-significant *p*-value does not provide evidence that there is zero or no effect at all [49,56]. Researchers should avoid using "magnitude based inference" [57,58] or "second generation *p*-values" [59] as *hypothesis tests* for equivalence, though they may be useful as *descriptive* statistics of the confidence interval. The primary problem with using these statistics as hypothesis tests is the higher false positive risk and dependence upon the sample size [57]. Instead, we encourage researchers to use well-established and valid tests of equivalence, superiority, or non-inferiority which have known error rates that do not change with sample size [53].

Second, do the accuracy of inferential statistics are typically dependent upon a number of assumptions. Most tests of statistical inference estimate parameters (means or correlation coefficients) and provide some inferential measure (e.g., t-statistic with an associated *p*-value). These "parametric" tests often assume that the distribution of the sample mean is roughly normal [60]. When these assumptions are seriously violated, such as with skewed data, either non-parametric statistical tests, data transformations, bootstrap or permutation methods may be necessary to make accurate inferences [61–63]. It is always good practice to assess the data (visually and statistically) to ensure that these assumptions are reasonably satisfied. Researchers should have contingency plans in place for their statistical analyses if these assumptions are violated. Without such plans in place, there can be numerous potential analyses to perform and researchers may be enticed to choose the result that is significant. A contingency plan for these violations and preregistration (discussed below) can help limit "data dredging" or "p-hacking" of the data to find significant results [6,8,45,47].

Overall, the process of deciding on the appropriate statistical tests should occur, at least partially, prior to the collection of data. There are a number of online statistical decision trees that can help in this decision making process [64], or alternatives can be found in biostatistics textbooks [65]. Within Table 1, we have included a number of common statistical approaches that may help in selecting the appropriate inferential statistical tests for a variety of experimental designs. All of these tests, including the non-parametric or robust options, have a variety of assumptions that should be reasonable satisfied in order for them to be useful, and researchers should be familiar with these assumptions prior to using any technique [19,60,62,66–69]. In any case, we strongly recommend that researchers consult a statistician to ensure an appropriate statistical analysis plan is generated prior to data collection.

A priori power analysis

Prior to data collection, it is important to ensure the experimental design will have adequate statistical power. It is important to remember that statistical power is a *conditional* probability. This means that the power of a statistical test can change based on the effect size the researcher is attempting to detect, the pre-determined significance level (alpha level), and the sample size. For instance, a study looking at changes from pre to post utilizing paired samples t-test, would have ~80% power to detect a 0.8 °C increase in core temperature, with a standard deviation of the change of 1.0 °C (Cohen's d_z = 0.8) with a sample size of 15 participants. However, this same study would be woefully underpowered (Power = 44%) to detect a 0.5 °C (SD = 1.0) increase in core temperature. Most researchers tend to use a power analysis to determine and justify the sample size for an experiment. Ensuring adequate power for an experiment is critical step in designing experiments. Underpowered studies, typically due to small sample sizes, are dangerous for two reasons: they lead to inflated effect sizes in the scientific literature [70] and can lead to the erroneous conclusion that there is no effect when one does exist (Type II error; Figure 2).

Power analysis has become an essential part of the research process for a number of reasons. Grant agencies and journals are increasingly requiring power analyses as a justification for samples sizes. Ensuring a study has adequate power is often an ethical concern. A study that is woefully underpowered to detect an effect is considered unethical and a waste of resources. In contrast, it would also be unethical to continue subjecting more participants to a research protocol, and use resources, when sufficient power can be achieved with a smaller sample size. Therefore, when conducting a power analysis, researchers are often looking for what is the number of participants or samples they will need in order to achieve the desired power.

In order to ensure a study's results are informative, power should be determined for the effect size the researchers would not want to miss detecting. In this case, it is useful to compare inferential statistics to a heat strain algorithm, which can be utilized to determine thermal strain during work in the heat [71]. In this case we would want an algorithm that is sensitive enough to detect a potentially dangerous work environment (i.e., a work load and environmental temperature high enough to cause heat illness), but not so sensitive that it deems any activity on a hot day

		Null hypothesis is ...	
		True	**False**
Judgement of the significance test	**Reject the null** (p < alpha)	**Type I error (Alpha)**	**Correct Decision AND True Positive (Power; 1-Beta)**
	Fail to reject the null (p ≥ alpha)	**Correct Decision AND True Negative (1-Alpha)**	**Type II Error (Beta)**

Figure 2. Chart demonstrating the statistical decisions based on null significance hypothesis testing.

dangerous. The same can be said about power analysis: we want to detect effects that are worthwhile but avoid declaring trivial effects significant. Therefore, researchers should determine what they consider the *smallest effect size of interest* (SESOI), or – in other words – the effect size a researcher does not want to miss. The SESOI can be based on measurement error (the minimal *detectable* difference; such as 0.05 kg on a weighing scale) [72] or the effect size that is large enough to be relevant (the minimal *important* difference; such as 0.12 kg on a weighing scale) [73,74] (please see the section on *Measurement Error*). It is up to the individual researcher to determine whether the minimal important difference or the minimal detectable difference should be used as the SESOI [73]. Further, researchers should carefully consider the manipulations utilized to produce the effects [75]. For example, observing a small change sweating sensitivity to a large loss in body water is much less impressive or important than a small change due to a small loss in body water.

The observed effect sizes reported in the literature and in pilot data *should not* be utilized for a priori power analysis because they do not reflect the SESOI [76]. The use of previously observed effect sizes will likely lead to an inadequate sample size estimation. Previous study effect size estimates should only be utilized when the sample size estimation can be adjusted for bias and uncertainty [77].

A pragmatic approach to power analysis, also known as a compromise power analysis, can be utilized when there are severe limitations on the maximum sample size. A pragmatic power analysis involves determining the alpha level and statistical power based on the sample size, the SESOI and the error probability ratio. The error probability ratio is the simply the beta error rate (inverse of power) divided by the alpha error rate (significance level). In essence, the error probability ratio is the ratio at which you are willing to have type II errors in comparison to type I errors. Typically, the error probability ratio is equal to four; considering most studies are designed to have 80% power (beta equal to 0.20) and an alpha of 0.05. Overall, this procedure should be used in cases where it is impossible to collect beyond a specific sample size due to costs, or when the sample size is entirely fixed (e.g., retrospective analysis of clinical records).

Researchers must be careful how they perform a power analysis, and be specific on the type of test statistic they will be utilizing. For example, it is common to see researchers report a power analysis for t-test, typically 1 pairwise comparison, when the experimental design utilizes a repeated measures ANOVA, with more than a single point group comparison. This is not an appropriate approach, and researchers should aim to have adequate power for the type of analysis, or analyses, they are performing (e.g., an interaction in a repeated measures ANOVA). The calculations necessary to perform a power analysis can be complicated for study designs that extend beyond a simple t-test. However, there are number of free [78–85] and commercially available [86] power analysis software options that can handle more advanced designs. In addition, there are other approaches to sample size planning, such as accuracy in parameter estimation that can be a useful alternatives to power analysis [87–89].

Preregister hypotheses, data collection, and analysis plans

It is important to consider, and report, what parts of a study are confirmatory or exploratory. Exploratory research is a vital part of the scientific process that allows researchers the degrees of freedom necessary to find interesting or vital information within the data. However, in order for a scientific theory to be falsifiable, prediction and confirmatory evidence are required. This is essential because science relies upon prediction to examine the validity or strength of a theory or hypothesis [9]. Moreover, significance testing is designed to be a confirmatory statistical test [9,90]. If researchers fail to distinguish between prediction and postdiction – or explaining the data after the fact – the scientific literature becomes distorted and researchers become overconfident in theories that are weaker than they appear. The mathematical psychologist Amos Tversky once eloquently summarized this problem.

"All too often, we find ourselves unable to predict what will happen; yet after the fact we explain what did happen with a great deal of confidence. This "ability" to explain that which we cannot predict, even in the absence of any additional information, represents an important, though subtle, flaw in our

reasoning. It leads us to believe that there is a less uncertain world than there actually is … [91]".

Researchers should aim to design and present studies in a way that separates the confirmatory (predictions) from the exploratory (postdictions) so that they do not give a false sense of confidence to themselves or other researchers. Preregistration is an effective method for separating the confirmatory from the exploratory. Overall, preregistration is the recording and committing to a study design, data analysis plan, and hypotheses prior to data collection or – in the case of using secondary data – without knowing the outcomes of an analysis [9]. Preregistrations can be easily committed on non-profit sites such as https://aspredicted.org/, through the Open Science Framework https://osf.io/, or, when clinical trials are involved, through the National Institute of Health https://clinicaltrials.gov/. We strongly encourage authors to use these preregistration resources considering the current evidence suggests they improve replicability [90], reduce poor statistical reporting [92], and improve the ability of other researchers to detect possible problems or errors [93].

Part 2: Data collection

During data collection, the statistics are often a secondary concern if a concern at all to most researchers. However, one serious statistical concern arises during data collection: the decision to stop data collection or to collect more data. Researchers do not want to collect more data when the current sample is sufficient, but researchers want to continue collecting data if there is still a good chance of detecting a worthwhile effect. However, repeatedly analysing the data violates the basic tenets of significance testing and can greatly increase the type I error rate rendering the *p*-value useless [94]. This is an unfortunately common practice [6,95], and recent peer reviewed research articles still contain statements like "we continuously increased the number of animals until statistical significance was reached" [96]. Instead researchers need to thoughtfully correct for this "optional stopping" of experiments. Sequential analysis provides a solution to the optional stopping problem [97].

Sequential analysis is particularly important in situations where data collection is highly expensive or prohibitive. In fact, scientists working for the United States military during World War II developed the methods for sequential analysis to ensure rapid deployment of new and helpful technologies. The US military thought the technique was so valuable that they deemed the material classified and would not allow the publication until after the war [98]. These sequential analysis techniques are essential for studies of medical therapies because they can provide early warnings of potential dangers or stop the trials early when it becomes clear a treatment is vastly superior [99]. There are numerous techniques for performing sequential analysis and below we have provided a quick guide (Table 3) to a few appropriate approaches for occupational and environmental physiology research.

Optional stopping and sequential analysis

Traditionally, type I error rate is conserved during sequential analysis of the data by adjusting the critical value, Z, thereby lowering the *p*-value threshold for significance. The original procedures for sequential analysis, the Pocock boundary and the O'Brien-Fleming procedure, are useful but only allow for a fixed number of interim analyses and equal number of observations between each analysis. For example, in a hypothetical experiment involving pre-post comparisons the researchers would have to set the fixed number of comparisons (let us assume that we want three "peeks" at the data), and the exact number of participants (let us assume that number is five participants). At 5, 10 and 15 participants data collection would be temporarily paused while the data are analysed to see if data collection needs to continue. Obviously, this is difficult because further data collection is delayed while the data analysis is performed and there is little flexibility on when data analysis can occur. If a researcher does not want to specify the number of interim analyses in advance, they can utilize a spending function which adjusts the alpha continually throughout the interim analyses [100,101].

If a researcher is interested in the accuracy of the estimated effect size then sequential analyses may make the estimated effect size more volatile [102,103]. In these cases researchers should utilize an approach termed accuracy in parameter

estimation (AIPE) [87,88,104]. Essentially, AIPE involves collecting data until the confidence interval is of an acceptable width [105]. This is a very useful tool if you want to avoid situations where you have a non-significant and non-equivalent effect (i.e., the confidence interval contains both zero and the SESOI). For example, imagine a study where you want to determine if pre-cooling improves cycling performance, but, from prior research [106], want to ensure that if there is a difference it is not less than 90 seconds. Therefore, we can set acceptable confidence interval width to 80 seconds. When we stop data collection because the confidence interval is of this width (80 seconds), if the confidence interval does include zero then it will exclude 90 seconds (therefore indicating equivalence), and if the confidence interval includes 90 then it will exclude zero.

The sequential analysis or AIPE procedures are promising but can be more computational demanding than traditional fixed sample size approaches. We highly recommend that readers fully understand these procedures before implementing them. There are numerous short tutorials [77,97,104,107] and textbooks [100,108] that are accessible to the average physiologists with little formal statistics training. In addition, there is free software and statistical packages that can help with sequential analyses [101,109,110]. There are some drawbacks to performing sequential analysis, and authors should be careful to report every detail about how the sequential analysis was performed (e.g., how many interim analyses were performed) [102]. In particular, researchers should be wary of the effect size estimates from a study that utilizes sequential analysis because the estimates are likely inflated [103]. We suggest that all researchers consider consulting a statistician for any advanced procedure such as sequential analysis.

Part 3: Data analysis

After the data are collected, the most time-consuming part of the statistical procedures begin. If parts 1 & 2 were completed correctly this portion of the statistics process is substantially easier and more time efficient. At this point, researchers must input the data into an appropriate system, "clean" the data by identifying potential errors, inspect the descriptive statistics, and produce the inferential statistics needed to test the hypotheses [111]. This process is tedious but essential for producing reproducible and replicable work. Special care must be taken to ensure the chosen analyses are appropriate for the data. Every statistical test has a number of assumptions – even robust and non-parametric options – that must be satisfied in order for the inferences to be worthwhile. Despite claims to the contrary, there are no "assumption free" statistical tests [112,p.201]. Furthermore, data analysis is about more than just dividing the results into significant and non-significant results, and researchers should thoroughly analyse their data to produce useful information regarding the size and uncertainty surrounding the observed effects.

Checking the data

Prior to any data analysis the arduous process of entering and checking the data must occur. Researchers should be aware of the database structures that are acceptable for the statistical software they are utilizing and ensure that the data are appropriately entered so that the data can be easily imported to that software. Further, the data should be labelled appropriately with clear variable names. For example, you may have many columns of rectal temperatures for experiments involving repeated measures. You may want to label the columns as "Trec" for rectal temperature followed by an underscore and a notation indicating the time point (e.g., "Trec_T1" is rectal temperature at time point 1). In these cases, researchers should have a codebook that details what each variable name means and what type of data are contained within that column so that anyone can understand the dataset.

When the data is being entered, researchers should also check for potential errors or problems with data. Catching errors early can save substantial time and grief later in the data analysis process. For example, while entering large amounts of data someone may accidentally misplace a decimal point and enter a value an order of magnitude greater or less than intended. Potential errors during data collection can also be caught during data entry by plotting the individual data points for the measure of interest. For example, a rectal temperature probe may exit the rectum during data collection and the recorded body

core temperature may suddenly drop below a physiologically plausible value, which will be clear when the data is visualized on a graph. These errors are very important to capture and if they are not caught and corrected, it will result in a flawed publication entering the scientific literature.

Violation of assumptions and robust statistics

Every statistical analysis has several assumptions, even the robust options discussed below, that need to be satisfied for the analysis to provide accurate results. Most physiologists utilize parametric tests that assume that the residuals from the model are normally distributed. In recent years, a number of statisticians have made it clear that normality should not be assumed without checks. When this assumption is not met it can be a source of erroneous or at least highly misleading results [62,68,113, p.6,114, p.36]. Moreover, the traditional methods for the detection of outliers also assume normality making it difficult to properly detect outliers [113,p.96]. Despite claims to the contrary, deviations from normality are *not* negligible once the sample size is large [113,p.8]. Instead, researchers should carefully investigate their own data and ensure the data is appropriate for the analysis they intend to implement.

Outliers

Outliers are simply data points that are inconsistent with the other observations in the data. These shift the measures of central tendency (mean) and the variance. In many cases, the outliers may be due to an error in data collection, but they may also be credible tails from a larger normal distribution or legitimate artefact. For example, an unusually high resting body core temperature could indicate a subclinical fever. Clearly erroneous data points should be removed prior to data analysis, and this should be reported.

In cases where a clear error cannot be identified, there are methods to aide in detecting potentially problematic outliers. A single outlier can be detected with the Grubb's method [115], and multiple outliers can be detected using either the Tietjen-Moore test [116] or the Generalized Extreme Studentized Deviate test [117]. However, all of these methods assume the data come from a normal distribution, and slight deviations from normality render these methods insufficient to detect outliers. Instead, Wilcox [118, p.35–38] recommends utilizing the median absolute deviation statistic, and visually inspecting all of the individual data points with a boxplot to aide in outlier detection.

If outliers are detected, researchers will need to decide about including these values in their statistical analyses. As suggested by Sainani [111], it may be useful to perform sensitivity analyses where outliers are included in the analysis then excluded from the analysis to see if the outlier substantially affects the statistical outcomes. If the outlier does not affect the conclusions then there is no harm in including these observations. On the other hand, if the outliers do affect the conclusions then both analyses (with and without) the outliers should be reported [119] so the influence of the outliers are clear. Regardless of how outliers are assessed, or if the data points are retained, the details of this process should be reported in the manuscript.

In longitudinal or repeated measures designs, the removal of outliers may lead to missing data points within a participant. For example, a participant may have had a rectal thermistor briefly fall out during data collection, and it would be necessary to remove that datapoint. However, if a repeated measures ANOVA is utilized to analyse the data then that entire participants' data is then removed from the analysis (listwise deletion). A researcher can replace the value using an imputation method (replacing the datapoint), or analyse the data using a generalized estimating equation or mixed model which can handle missing data without dropping the participant entirely. There is not a single default method for dealing with missing data, and researchers should carefully consider all available options [120].

Data transformation

Sometimes the data is skewed (Figure 3), possibly due to outliers, and it may be advisable to transform the data in order to satisfy the assumption of normality. In fact, some empirical investigations would indicate that data transformations are sometimes superior in normalizing the data compared to eliminating outliers from the data [121]. There are numerous data transformations and researchers should be careful in which type they

utilize [61,112,p.192]. While the log transformation is the most common type of transformation it is often not the most appropriate [61]. There are general power transformation procedures that can help simplify the selection process such as Tukey's transformation ladder [122] and the Box-Cox family of transformations [123]. The type of transformation should be chosen based on how well it normalizes the data *not* on whether the transformation provides a significant result (Figure 3).

In small samples, it can be exceedingly difficult to determine if there are deviations from normality or if a transformation is effective in normalizing the data. In fact, data transformations are often criticized because they fail to normalize the data [124]. As we can see from residuals to Figure 3, the while the transformed data (3D) appears to be slightly more normal than the raw data (3C) it is not perfectly normal (3D) following the transformation because of the heavy tails of the distribution. In addition, the interpretation of the data becomes tricky once the data transformation occurs because now it is an analysis of the geometric means not the original arithmetic means. Despite these drawbacks, the transformation of data can sometimes be useful, but the process of selecting a transformation should be carefully considered prior to implementation [112,p.191–202].

Robust alternatives

Sometimes data transformations and outlier elimination methods are ineffective at normalizing the data. Furthermore, in small samples, tests of homogeneity of variance (discussed below) and normality are underpowered to detect violations of these assumptions. This makes it difficult to determine if these assumptions are violated or even if transformations or outlier exclusion are helpful. Occupational and environmental physiologists may often find themselves in such a position where the use of parametric tests – such as the ANOVA or t-test – may be dubious. Therefore, a researcher may need to use non-parametric tests, which do not assume normality, or robust approaches, which are not greatly affected by changes to the underlying distribution [113].

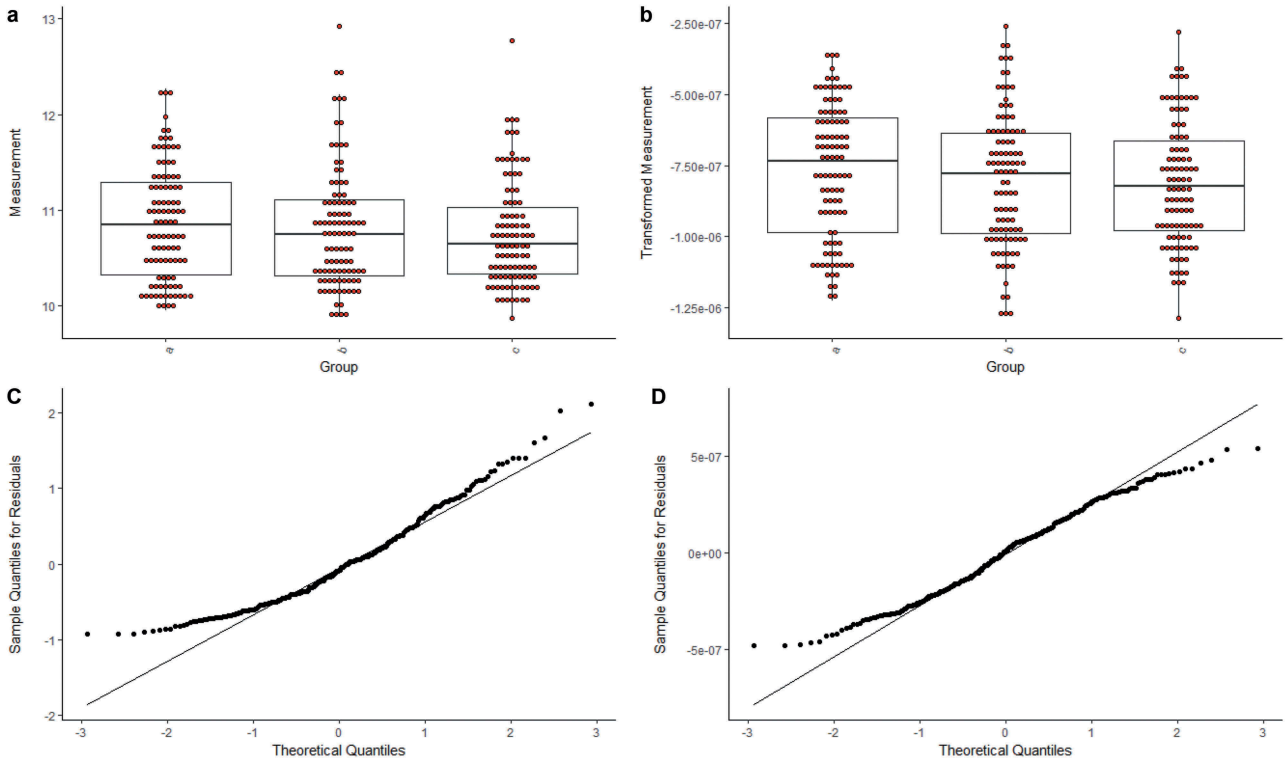


Figure 3. Demonstration of skewed (a) and transformed data with three groups. In addition, visualization of the residuals for the skewed (c) and transformed (d) data.

According to Good [114] and Efron [63], permutation and bootstrapping methods have fewer assumptions than traditional parametric tests and therefore offer a distinct advantage. In particular, with small sample sizes it is entirely impossible to tell if the assumptions of normality are met, and permutation methods are the preferred default approach for simple t-test and one-way ANOVA type designs [62]. In addition, when outliers are a concern permutation methods, again, are more robust than standard parametric options [114,p.198–200].

Other robust methods for statistical testing and estimation can be utilized when the data are not normally distributed, and data transformations are not helpful. Instead, trimmed means, Winsorized means, or maximum likelihood type estimators (M-estimators) are all viable alternatives. Unlike permutation methods, these robust statistics can be easily applied to repeated measures or factorial type experimental designs. Further, these measures, unlike permutation tests, are fairly accurate when the distributions are asymmetric [112,p.201].

Permutation and other robust methods have been around for a long time, but only recently have become feasible due to massive improvements in computing power. Robust statistical methods are now available, for free, in the R programming language through the WRS2, bootES, and jmuOutlier packages (and numerous other packages) [125–127]. Many robust procedures are also available within the commercial software StatXact (Cytel; Cambridge, MA). Some of these methods have been implemented for point-and-click usage within the free statistical software Jamovi under the "Walrus" module. These robust methods are particularly useful when the sample size is small since it may difficult to detect violations and the robust tests tend to be more powerful [118,128]. However, it is up to the individual researcher to decide what techniques are appropriate for their research. We highly recommend that researchers decide prior to data collection what statistical techniques they should utilize or at least have contingency plans in place in order to avoid ad hoc "p-hacking" practices [6].

Beyond normality: What else should I consider?

Most statistical tests also have a number of other assumptions that should be considered. We will discuss some these assumptions below, but these assumptions are discussed in-depth elsewhere [112,p.166–204]. We encourage readers to understand all the underlying assumptions for any statistical tests they utilize.

First, most tests assume the data is homoscedastic, or that the variance is stable across groups or the predictor variables. This is commonly called the assumption of homogeneity of variance. This means that the groups being compared come from populations with the same variance, or – in regression – the variance for the response variable is stable across the predictor variables (Figure 4). Like the assumption of normality, there are tests, such as Levene's or Hartley's F_{max}, to test for the assumption of homoscedasticity, but in many cases these tests will be underpowered to detect such a violation. These assumptions can also be evaluated through visual inspection of the model residuals (Figure 4(b,d)). If such violations are detected then a power transformation (e.g., Box-Cox or Tukey's Ladder transformations) can reduce heteroscedasticity, or robust regression can be utilized (see above).

Second, when utilizing regression it is assumed that the relationship between variables is linear. This means that for every incremental change in the predictor variable there is a proportional increase in the response variable (Figure 5(a)). There are many cases in physiology where this assumption should not be assumed. For example, the relationship between plasma vasopressin and plasma osmolality changes along the physiological range and is better described using segmented regression [129] or fitting the linear regression with polynomial (i.e., quadratic, cubic, quartic, etc.) or reciprocal (i.e., 1/predictor) term. The assumption of linearity can be easily assessed using the residual plots (Figure 5(b)). If the relationship is assumed to be not be linear then more advanced regression techniques, such as polynomial regression may be useful [130,p.520]. When a quadratic term (x^2) is added to the model (Figure 5(c)), the residuals are approximately randomly distributed (Figure 5(d)).

Third, most inferential statistics assume the data are independent or that outcome of one observation is not dependent upon another. Obviously, this is not the case for participants within a repeated measures design, but it still holds for comparisons between participants (e.g., observations within one

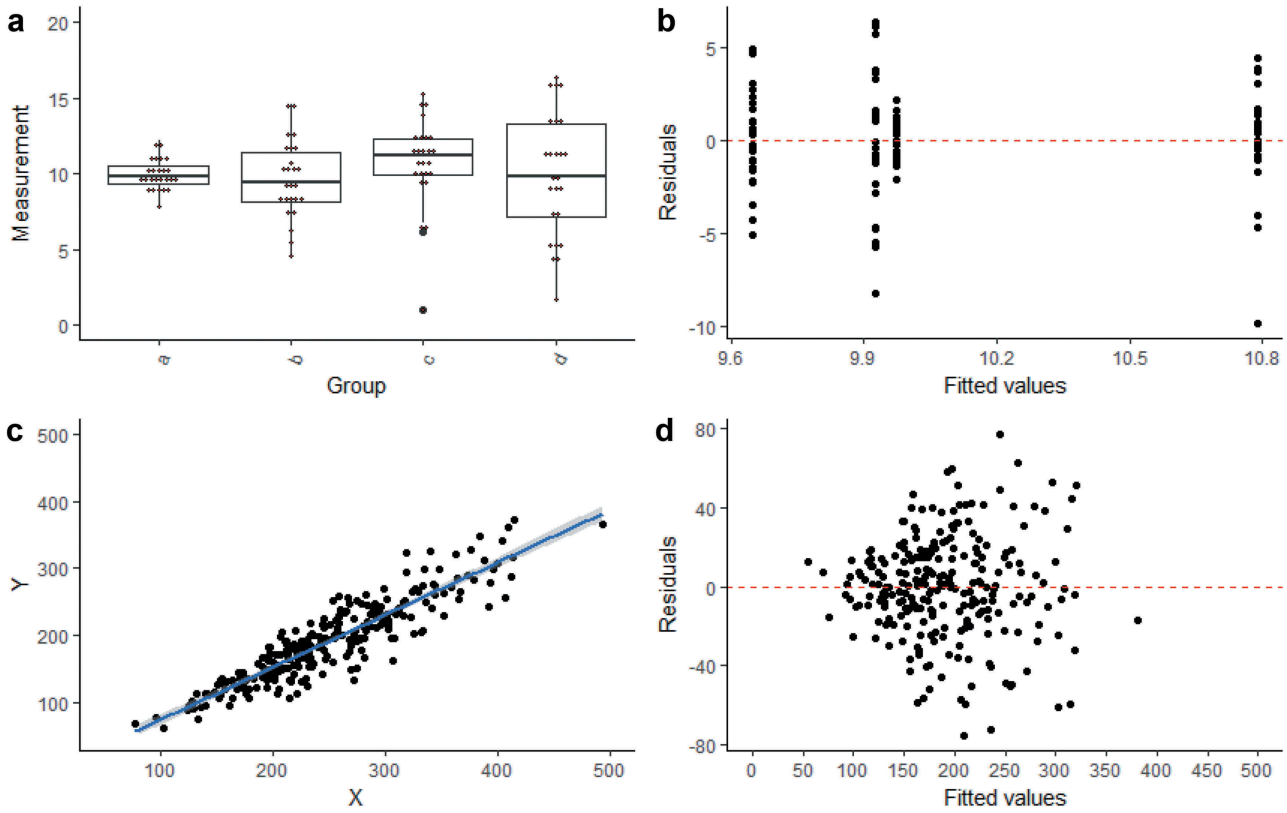


Figure 4. Heteroscedascity when comparing groups (a) and in regression (b). Violations of the assumption of homogeneity of variance can be visually diagnosed with residual plots for either categorical (c) or continuous (d) variables.

participant should be independent from other participants). This includes designs where multiple observations within a participant are utilized to predict an outcome of interest (e.g., core body temperature) [131]. In these cases, using multi-level or hierarchical models may be a useful alternative when the assumption of independence is violated (see Table 1) [132].

Fourth, when utilizing an ANOVA, the problem of multiple comparisons arises. When using NHST, the more statistical tests you perform the more likely you are to observe a significant result. This becomes particularly problematic when there are multiple groups and therefore multiple statistical tests that can increase the type I error rate. There are procedures that correct for these multiple comparisons by correcting the test statistics themselves (e.g., Tukey-Kramer, Newman-Keuls, and Scheffe), or by adjusting the *p*-value (e.g., Bonferroni, and false discovery rate procedures). When utilizing repeated measures there are limited options because powerful procedures, such as Tukey or Scheffe, do not control type I error in these situations. Further, some commonly reported procedures, such as Newman-Keuls, do not control type I error rate regardless of the study design [133]. There is no single "best" post-hoc comparison correction. Researchers should understand the advantages and potential pitfalls each procedure and then decide if they are appropriate for the experimental design at hand [133].

Calculating and interpreting effect sizes

An effect size is defined as, "a quantitative reflection of the magnitude of some phenomenon" [134]. In the physiological sciences, unstandardized effect sizes (i.e., mean differences) are very common, and are useful when the raw differences are interpretable. For example, we can easily express changes in core temperature during heat stress as the mean difference in Celsius or Fahrenheit with a 95% confidence interval. However, standardized effect sizes are useful when the measured values are arbitrary units (e.g., Likert-type scales) or in meta-analysis when comparing effects measured on different scales or devices. Standardized effect sizes are often calculated

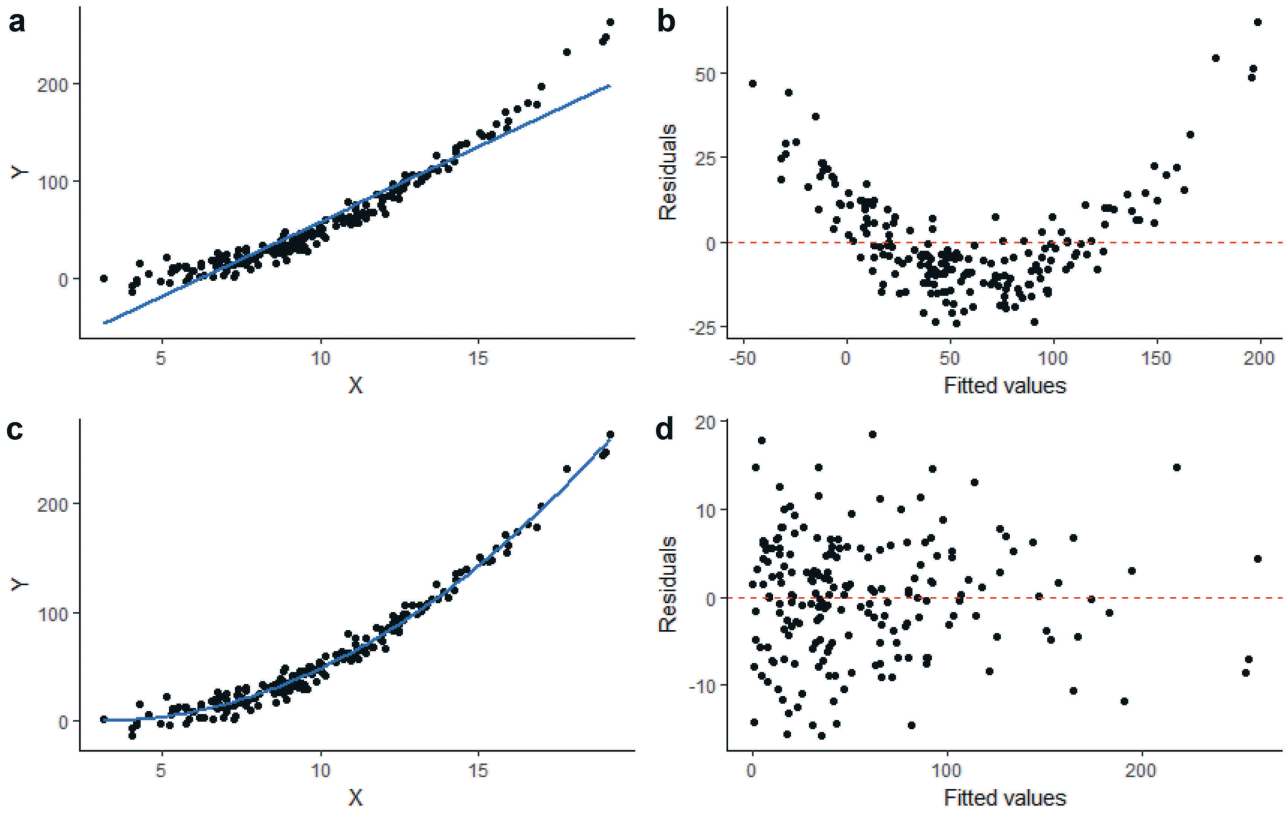


Figure 5. Plots of a curvilinear (quadratic) relationship with data fit using a (a) linear (first-order) model with (b) the associated residuals, and fit with a (c) linear polynomial (second-order) model and (d) the associated residuals.

as the mean difference divided by the standard deviation (or at least some variation of this).

It is important to separate what is statistically significant from what is practically significant or relevant. Effects sizes allow researchers to interpret the magnitude of the effect thereby providing the practical significance of the results. For example, whole number ratings from a commonly used thermal sensation scale [135], when analysed, may produce mathematical fractions of a rating which result in statistically significant differences with large sample size but are of questionable value. Imagine, for example, a study observing average thermal sensation ratings of 6.5 versus 6.1 where the differences are statistically significant ($P < 0.05$). Both ratings are indistinguishably "warm" on the scale and both are associated with ~2^{0}C increases in mean skin temperature above "slightly warm" but below "hot" [135]. In other words, the quantitative difference is smaller than the smallest categorical difference, thereby being below the minimal detectable difference. The use of standardized effect sizes and confidence intervals would improve interpretation greatly in this example.

Interpretation of the effect sizes can become problematic. Cohen's effect sizes for the social sciences are often cited, but effect size interpretations observed in environmental and occupational research may be entirely different. This has been discussed and quantified in depth for a variety of topics in physiology [136–139]. The most important thing to remember is that effect size interpretation is specific to the outcome of interest. For example, a change in Likert scale response of 0.5 SD is typically considered a sizeable effect whereas a 0.5 SD change in many physiological parameters (e.g., rectal temperature) may be rather small or even inconsequential. Simply calculating the effect sizes and then interpreting based on Cohen's scales (e.g., an effect is large because $d = 0.8$) is an inadequate and likely misleading approach to interpreting effect sizes in environmental and occupational physiology.

Researchers can instead decide what a relevant change or effect size is prior to data collection, and interpret accordingly after data analysis is completed. A default of 0.5 standard deviation difference [51] is a fine default for a clinically meaningful difference in the absence of other information.

However, we encourage researchers to establish a SESOI prior to data collection based on empirical evidence.

Researchers should also be careful in interpreting studies with small samples because effect sizes are likely to be exaggerated [103,140]. Most researchers should, by default, adjust for bias in small samples by applying a Hedges correction to standardized effect sizes [141]. Even with this correction, or others [125], the effect size estimate may be imprecise and researchers should express the uncertainty about the effect sizes by providing the confidence interval around the effect size estimate. Calculating confidence intervals requires the use of noncentral *t*-distributions for which there is no generic formula [142]. Instead, we highly recommend that researchers calculate robust confidence intervals through bootstrapping procedures when possible [125,142]. Bootstrap confidence intervals can be generated in R [125,143], SPSS (BOOTSTRAP Command, https://www.ibm.com/support/knowledgecenter/SSLVMB_24.0.0/spss/bootstrapping/idh_idd_bootstrap.html), and in SAS using macros (http://support.sas.com/kb/24/982.html).

The calculation of the appropriate effect sizes and corresponding confidence intervals can be difficult and time consuming. In particular, standardized effect sizes can be challenging to calculate considering the most common statistical software programs, such as SAS or SPSS, do not include options for calculating effect sizes as default options. However, open source options such as Jamovi and JASP include standardized and unstandardized effect size as default options when utilizing ANOVAs or t-tests. Furthermore, a variety of standardized effect sizes, and the confidence intervals, can be calculated in freely available spreadsheets [50,144]. An even wider variety of effect sizes and confidence intervals can be calculated in R through the packages compute.es [145] and bootES [125].

Confidence intervals and demonstrating uncertainty

One study is never the last word on a topic and cannot "prove" a phenomenon exists. Therefore, it is important to show uncertainty surrounding an effect. This can be partially accomplished by calculating the confidence interval around estimates. Statisticians often lament that confidence intervals are not reported for effect sizes or other estimates because researchers are too embarrassed by their width and are unlikely to admit to such a high level of uncertainty [146]. We encourage authors to overcome this inclination and be open about uncertainty.

Overall, a confidence interval shows the plausible values for the estimate. Contrary to popular belief, confidence intervals do not indicate the probability that the true estimate is within the interval [147]. For example, a 90% confidence interval demonstrates that, if an infinite number of experiments were to replicate the original results, then 90% of the confidence intervals will include the actual population mean.

Confidence intervals have a number of assumptions that should be met in order be an accurate representation. However, modern computing methods make alternative confidence interval estimates, through bootstrapping methods, easy to calculate for effect sizes [125]. When the assumptions of traditional confidence intervals cannot be reasonably met then alternative robust measures can and should be utilized [113,p.112].

Part 4: Dissemination of results

> "Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results. When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid sole reliance on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information. … Give numbers of observations. … References for study design and statistical methods should be to standard works (with pages stated) when possible rather than to papers where designs or methods were originally reported. Specify any general-use computer programs used." – International Committee of Medical Journal Editors [148].

When writing up the results of study, the statistical methods should be described in enough detail to allow readers to *verify* the results and understand if the reported analyses were appropriate. Authors should avoid presenting just a *p*-value. Instead, authors should present other relevant information such as an effect size (e.g., mean difference,

correlation, or Cohen's d), test statistics (e.g., t-value or F-ratio) and associated degrees of freedom – at least for the primary outcome(s) of interest. However, avoid the temptation to include analysis details for every analysis or outcome measured in the study (i.e., p-clutter) [149,p.117–141]. Also, when a *p*-value is presented the precise *p*-value (two or three decimal places) should be reported not the level of significance (e.g., $p = .023$ not $p < .05$) [150]. However, *p*-values less than .001 should be reported as $p < .001$ [151,p.114]. All of this information is necessary to verify and evaluate the validity of a typical statistical analysis.

Currently, reporting statistics and data in the physiological sciences is very poor. In many cases (21%), the sample sizes per group/condition are not reported or only a range is provided [119]. As we have detailed in this review, a priori power analysis is an essential part of designing experiments, but only 3% of studies in cardiovascular physiology report an a priori power analysis [119]. In addition, ~20% of physiology research articles reported assumptions, such as normality, were checked or verified [119,152].

As a field, we can certainly do better than the current status quo. Below we have detailed several simple steps researchers can take to improve how their studies and experiments are reported. There are a number of guidelines for reporting data, and we encourage researchers to seek out the standards specific to their field of study. The EQUATOR Network (https://www.equator-network.org/) has made this process easy by aggregated all current reporting guidelines. We encourage readers to read the guidelines specific to their research, and report their results in line with these recommendations.

Open science & data sharing

Technology has unlocked the possibilities for scientific outreach and collaboration. Open science is the process of using these capabilities to make the research process more transparent, accessible, and efficient to other researchers and the public. Utilizing these capabilities is critical because science relies upon the ability of others to critically evaluate a study's evidence, reproduce the results and potentially replicate the result when necessary. When statistics are inadequately reported it makes it difficult to verify and reproduce scientific findings [153]. In addition, lack of openness and data sharing make it difficult for future researchers to follow-up, replicate, or reproduce your work thereby limiting the usefulness of a study. When reporting the results of an experiment or study researchers should go beyond simply writing the summary statistics and significance. Grant agencies, such as the National Institute of Health (https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#goals) and National Science Foundation (https://www.nsf.gov/bfa/dias/policy/dmp.jsp), are increasing requiring or strongly encouraging data sharing and transparency. The default practice should be to make the data, and the analyses of the data, publicly available.

The evidence base researchers build with their statistics are dependent upon the data that has been collected and the methods used to analyse that data. When the data and the methods to analyse the data are kept hidden so too are critical mistakes and oversights. As scientists, we are in search of the truth, and no matter how uncomfortable or embarrassing we should want our errors to be identified in an efficient and clear manner so that the scientific record can be corrected. In some situations, providing open data is not possible due to privacy or other ethics concerns. Researchers should be careful to safeguard the privacy and protect confidential, or proprietary, information. However, the evidence is clear that, when possible, data and the code used to analyse the data should be shared to facilitate future analyses and aide potential error detection [154].

The more advanced, complicated, or unusual your statistical analysis the more you will need to report in order for an adequate review of your statistical procedures. In many cases, it may be vital to show how exactly the analyses were performed by providing the data and the computer code/scripts. However, journals often have word limits and very long methodology sections can detract from an articles narrative structure. Therefore, we encourage authors to report additional details in a supplementary material section or host the information on separate sites such Open Science Framework of FigShare. Overall, open data and scripts can allow reviewers and readers to verify your findings without distracting from the narrative of the research article.

Data sharing should be the default practice for the majority of occupational and environmental physiologists in order to preserve scholarly record [154]. In addition to error detection, data sharing and open science practices can accelerate the research process by allowing for "multiverse" analysis of datasets [155]. However, researchers should take caution when preparing data for release, and ensure that participant anonymity is protected.

Data visualization

For those unaware, data visualization refers to the process of creating figures or graphs. Proper data visualization is critical because figures are the main way to convey key findings. As Weissgerber et al. warn, "pretty is not necessarily perfect" when it comes to data visualization [2]. Instead, the data should be reported in a way that allows the reader to critically evaluate the data. By far the most common type of data visualization is the bar graph (~85% of all graphs), but this type of visualization is most often used inappropriately [152]. Bar graphs should be used for categorical or count data; not continuous distributions [152]. Also, the standard error of the mean (i.e., SEM) is a commonly reported, but typically considered a misleading descriptive statistic [150]. Instead, we encourage authors to report, and visualize, the standard deviation or an appropriate (typically 90–99%) confidence interval around the mean depending on the intent of the visualization [150].

Regardless of the graph type used, the ratio of the size of any effect shown on any graph to the size of the effect in the data results should always be between 0.95 and 1.05 [156,p.57]. For example, Figure 6(a) illustrates a hypothetical 30% difference in reported heat illnesses between two calendar years, whereby the graphical to data ratio is 1.0 (both are 30%). In Figure 6(b), the graphical difference is 60% but the data difference remains 30%. The ratio is 2.0 because the y-axis scale has been altered. In order to ensure that numbers as physically measured on a graph are proportional to the quantities represented, the y-axis should include the realistic range of data that prevents "Lie Factor" distortion. In the case of physiological measures like heart rate or body core temperature, the scale should include the meaningful physiological range. Clear graphics labelling or the alternative use of tabular data may also be used to prevent distortion when "Lie Factor" rescaling is problematic [156,p.57].

There are number of other poor visualization practices that have been discussed previously but are beyond the scope of this review [152,157]. Below we have included a number of quick guides to creating data visualizations for different types of common experimental designs and situations. The code to re-produce for all the figures is provided in the supplementary materials (note: all code written in R). Further, for those unfamiliar with R, previous publications have produced Excel spreadsheets [144,152] and online applications [157] that can be helpful in creating data visualizations.

Independent samples

In studies with small samples ($n < 15$), it is best practice to provide univariate scatterplots so that individual data points are visualized. When just the summary statistics are presented (e.g., mean and standard deviation) problems with the data may remain hidden. Figures with individual data points allow the reader to assess the distribution for abnormalities. This is important because the individual data points may indicate if parametric statistics are appropriate [152]. Summary statistics, such as the mean and standard deviation (Figure 7(a)) or median and interquartile range (Figure 7(b)) can be superimposed upon the data points. In cases where parametric statistics (t-test) are utilized the summary statistics should be represented by the mean with errors bars that show the variability of the sample (e.g., standard deviation) (Figure 7(a)). When utilizing non-parametric or robust tests the median and interquartile range (boxplot) should be presented instead (Figure 7(b)).

Paired samples

Similar to studies with independent samples, authors should create data visualizations with the individual data points. However, with paired samples the statistic of interest is the *change* within each subject (e.g., *within*-subjects design). Therefore, it is important to show the individual changes from sample-to-sample (Figure 8(a)) and by showing the distribution of the difference values (Figure 8(b)). Again, the summary statistics that are

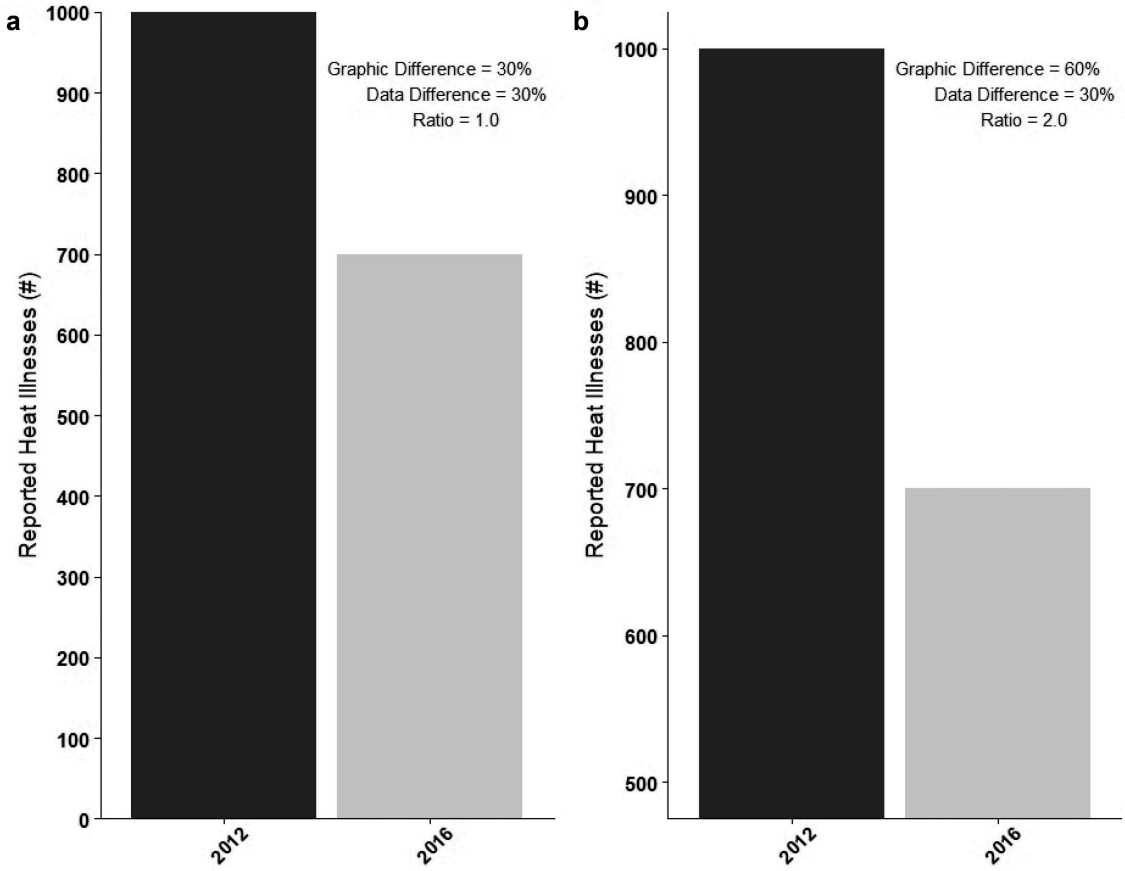


Figure 6. Theoretical comparison of reported heat illnesses using appropriate (a) and inappropriate (b) y-axis scaling.

presented should be based upon the type (parametric versus non-parametric) of statistical tests utilized to analyse the data.

Repeated measures

In repeated measures designs, it can become quite difficult to visualize the differences across many time points and between groups. Unlike a simple paired comparisons design, we cannot simply show all the individual changes. For example, let us imagine a study where we are comparing changes in core temperature in an industrial environment with (experimental; n = 6) and without (control, n = 6) a new modality to aid in cooling individuals. If the figure were to include all the individual change lines the figure would become far too cluttered. Instead, authors can still provide the individual data points with multiple summary statistics (boxplot and mean with confidence interval) superimposed with lines connecting the summary statistics to indicate the connection between repeated measures. This type of visualization may seem difficult to create, but can be easily generated in R (Figure 9).

While these line graphs are helpful visualisations they cannot demonstrate within-subject changes between time points. Interactive figures, wherein individual data can be modelled across time points, can be produced as supplementary material within a manuscript [157]. These interactive figures can be very helpful for readers exploring individual changes within the dataset but keeps the static figures within a manuscript organized and easy to understand.

Large samples

For larger samples, the assumptions regarding the distribution of the data can be inspected and researchers can proceed with parametric tests if these assumptions are not seriously violated. Data visualization with large samples can be overly complicated and busy if all the data points are presented. Therefore, in these situations it is advisable to present the summary statistics along with some visualization of the

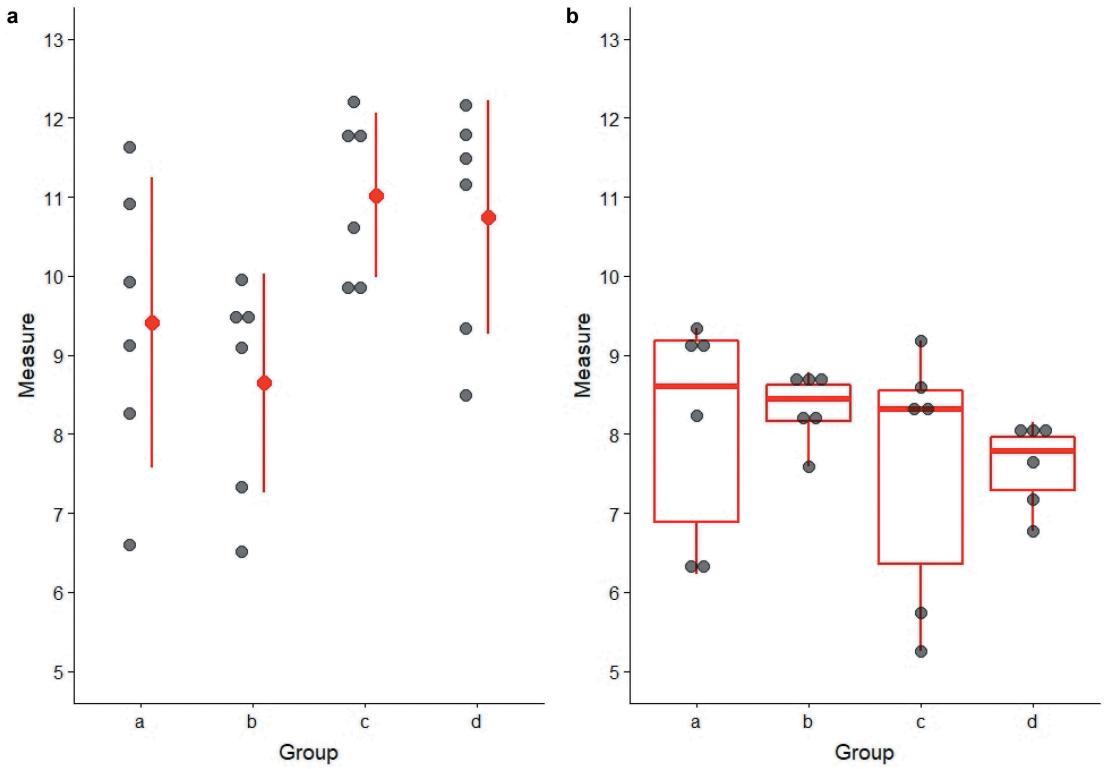


Figure 7. Demonstration of acceptable visualizations for simple independent group comparisons. The data can be visualized with dots of the individual observations with summary statistics of (a) the mean and standard deviation or (b) boxplot with the median, and interquartile range.

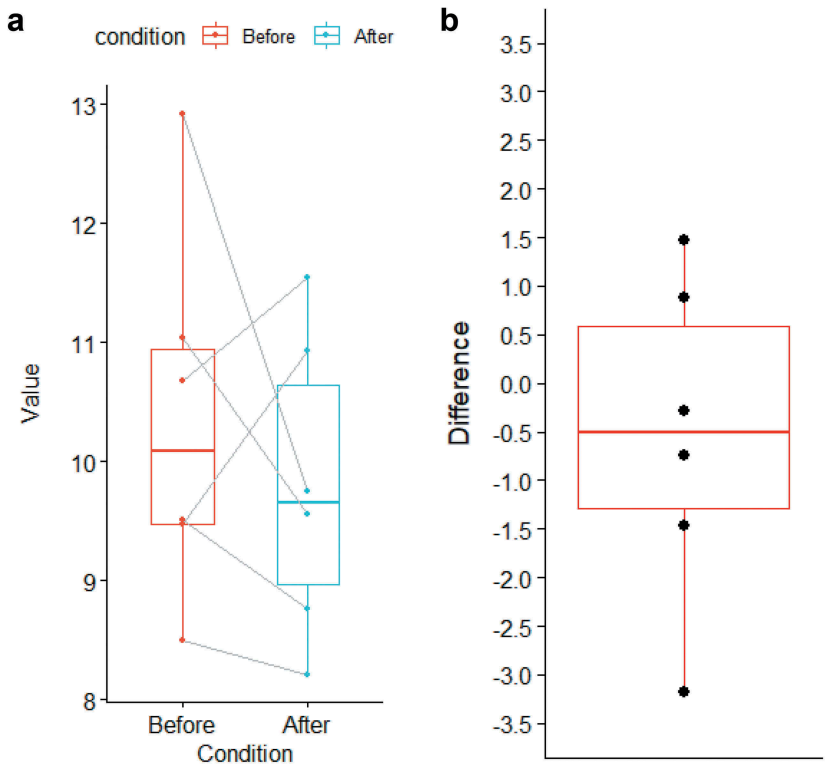


Figure 8. Visualization of paired samples comparisons. This type of data can be visualized by showing both samples with individual slopes for each participant (a) and by providing the individual differences (b) with a box plot to display summary statistics.

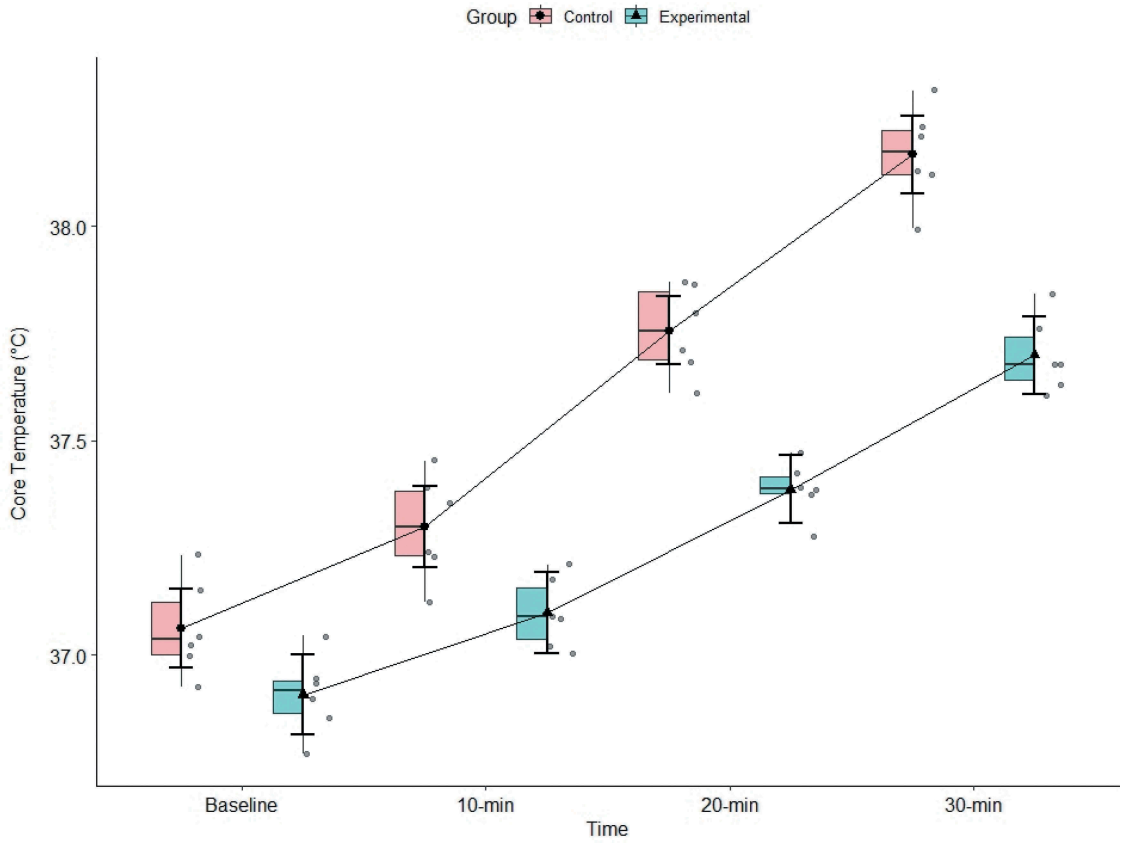


Figure 9. Demonstration of an adequate visualization for repeated measures design. Each time point has the mean with 95% confidence intervals (black), a boxplot displaying the median and interquartile range (pink and turquoise), and individual data points (grey).

distribution rather than the data points. This can be accomplished with violin plots (Figure 10(a,b)).

Reporting uncertainty or ambiguous results

Researchers should not be afraid to state that the data is uninformative or at least uncertain. In cases with small sample sizes [103,140,158], it may be too hasty to conclude that there is clear evidence for the presence or absence of an effect. Furthermore, effect sizes at small sample sizes may be highly volatile even when no effect exists. As Figure 11 demonstrates large effect sizes (Hedges' $g > 0.8$) are common when the sample size is small even when the null (no difference) hypothesis is true. This is why we encourage authors to present the confidence intervals around effect sizes to help convey the level of uncertainty around any single estimate.

Researchers need to embrace uncertainty, and make it clear to the reader when the results are ambiguous or tentative. This is particularly important because one single study is almost never enough to provide definitive evidence of an effect or phenomena. When writing research articles, authors should avoid creating a sense of certainty, and instead focus onproviding the information necessary to replicate and build upon their work available.

Conclusion

Statistics should not be a mindless, cookbook-like procedure that is completed at the conclusion of a study. In order to produce useful statistical information, researchers need to consider statistics as a process that starts far before data are even collected. Mindless and ritualistic use of statistics has created numerous replicability and reproducibility problems within a variety of scientific fields [3]. No magical or special statistical procedures exist that safeguard against statistical pitfalls and mistakes. Every step of the research process requires careful consideration. In this review, we have raised a number of points that every environmental and occupational physiologist should consider during the process of designing an

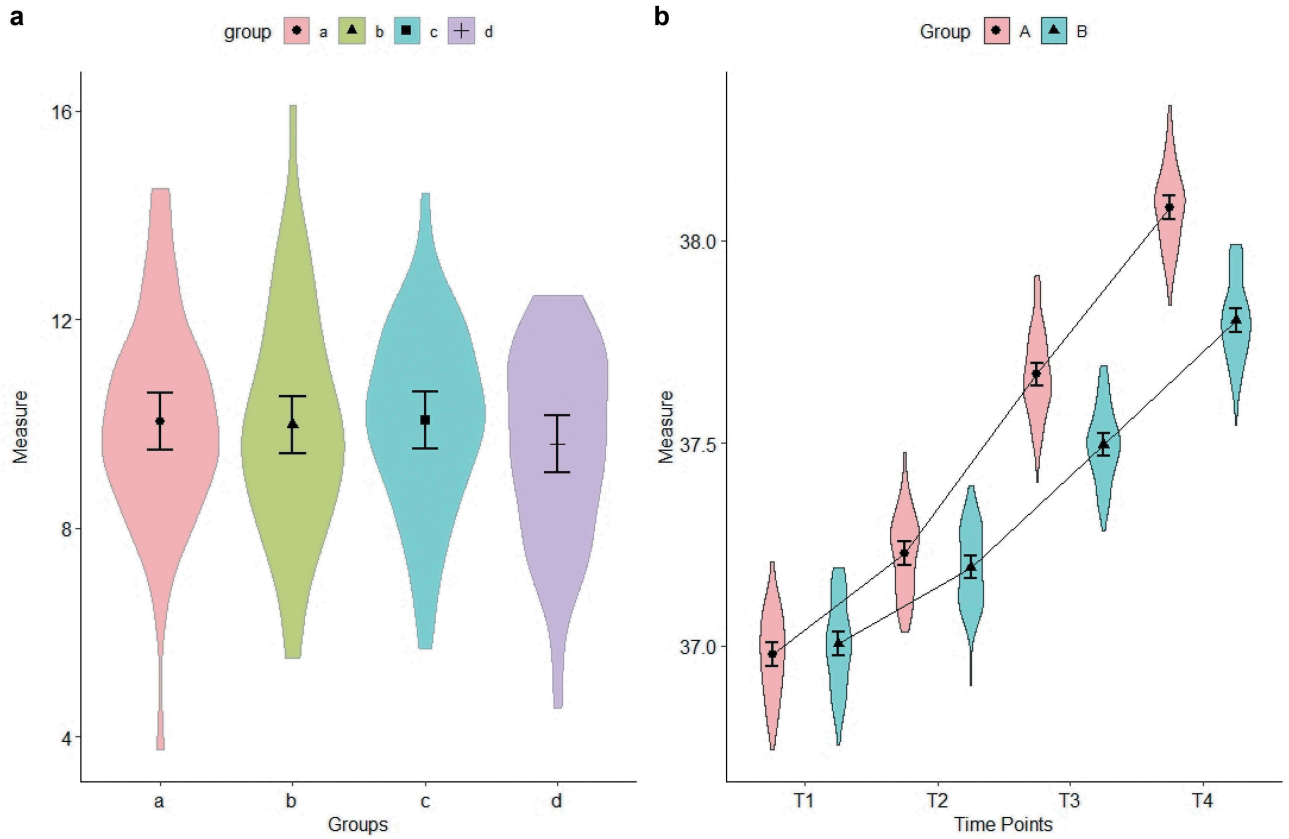


Figure 10. Visualization for large samples, in this case n = 50 per group. Violin plots are utilized to show the distribution of the data. Independent samples (a) can be simply visualized with mean and 95% confidence interval with a violin plot surrounding the summary statistics. Studies with repeated measures (b) can be visualized in a similar way with the only exception being a trace line connecting the time points.

experiment (Table 2), collecting data (Table 3), analysing the data (Table 4), and publishing their results (Table 5). We encourage researchers to consider these points and then decide on which procedures or approach is appropriate for the case at hand. All researchers should take the time to learn the basic concepts underlying the statistics they utilize. Beyond this, when performing any statistical analysis researchers should reflect upon and examine a method's capabilities and shortcomings. Researchers should generally understand when certain statistical tests are appropriate, and recognize when they need to consult a statistician for complex analyses.

Further reading

While the reference list contains many good areas for further reading, we have decided to highlight a few texts that are very helpful for understanding appropriate statistical practices. There are a number of statistical misconceptions that persist among researchers, which are documented by Gerd Gigerenzer in "Statistical Rituals: The Replication Delusion and How We Got There" and in Schuler Huck's book "Statistical Misconceptions". For a better understanding of the history and philosophy of inferential statistics, we highly suggest reading Zoltan Dienes' "Understanding psychology as a science" as well as David Salsburg's "The Lady Tasting Tea". Deborah Mayo has also tackled the recent debates regarding statistical inference in her new book "Statistical Inference as Severe Testing". This review also focused on a "frequentist" perspective, and for different "Bayesian" perspective we encourage readers to read Richard McElreath's "Statistical Rethinking: A Bayesian Course with Examples in R and Stan". For an introduction to robust statistical methods consider reading Rand Wilcox's "Fundamentals of Modern

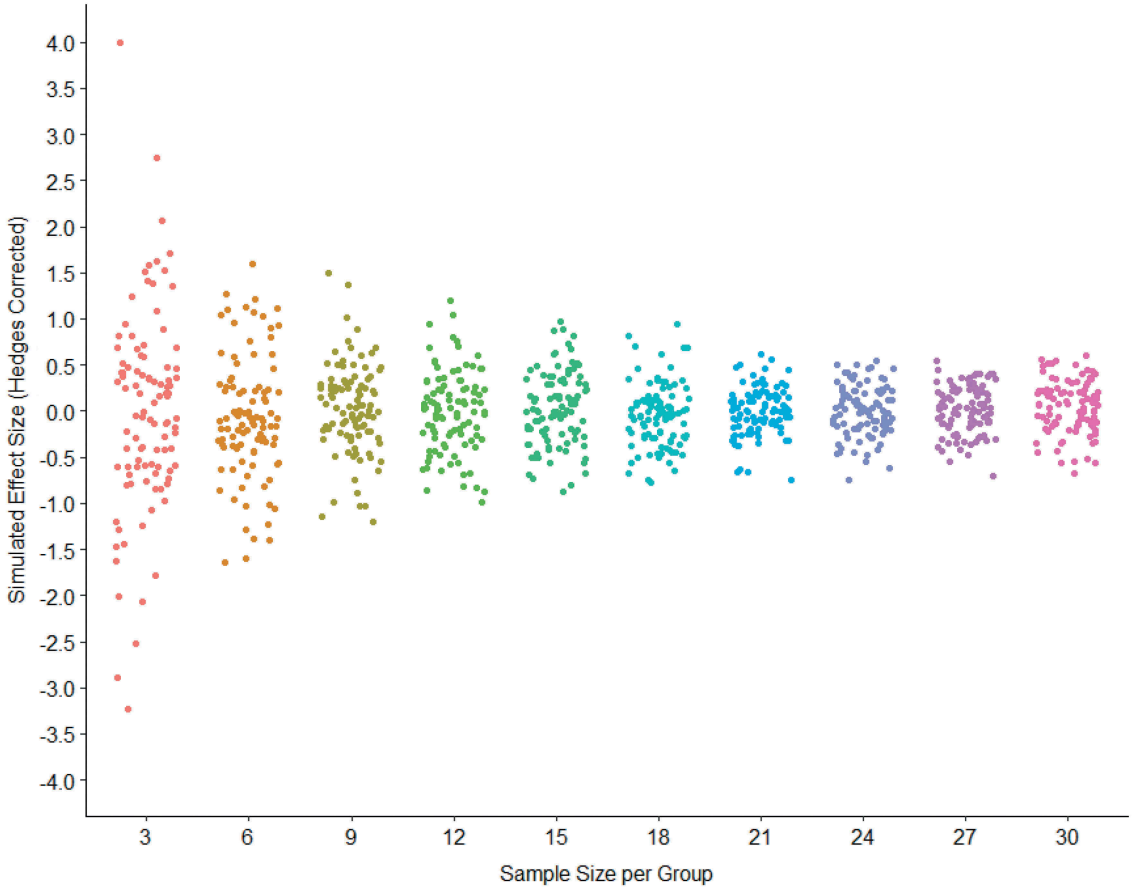


Figure 11. Effect size sizes under varying samples sizes (x-axis) when the null hypothesis is true (true effect size = 0). Larger effect sizes, upwards of g = 4.0, are possible when the sample size is small (n = 3 per group). Effect sizes were produced from 100 simulations per sample size (see supplementary material for code).

Statistical Methods: Substantially Improving Power and Accuracy", and Philip Good's "Permutation, Parametric, and Bootstrap Tests of Hypotheses".

Acknowledgments

The authors would like to thank Megan Rosa-Caldwell, Whitley Atkins, and Katie Stephenson-Brown for their constructive criticism and proofreading of this manuscript. The opinions or assertions contained herein are the private views of the authors and should not be construed as official or reflecting the views of the Army or the Department of Defense. Any citations of commercial products, organizations, and trade names in this report do not constitute an official Department of the Army endorsement of approval of the products or services of these organizations. Approved for public release: distribution unlimited.

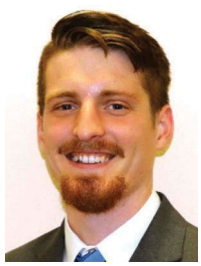
Abbreviations

AIPE	Accuracy in parameter estimation
ANOVA	Analysis of variance
CV	Coefficient of variation
NHST	Null hypothesis significance testing
TOST	Two one-sided tests
SESOI	Smallest effect size of interest

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors



Aaron R. Caldwell, Ph.D., received his PhD in Health, Sport and Exercise Science at the University of Arkansas while also completing a graduate certificate in Statistics and Research Methods. He is now starting a postdoctoral fellowship at the US Army Research Institute of Environmental Medicine. Aaron also serves as a board member for the Society of Transparency, Openness, and

Replication in Kinesiology (STORK), and the Chair for the preprint server SportRχiv. His current research efforts are focused on the application of statistics within physiology.

Samuel N. Cheuvront, Ph.D., R.D., FACSM, is the Deputy Chief of the Biophysics & Biomedical Modeling Division and Research Physiologist at the United States Army Research Institute of Environmental Medicine (USARIEM) in Natick, Massachusetts. His research interests include the broad study of environmental and nutritional factors influencing exercise performance and health with emphasis in hydration, heat stress, and modeling of sweat losses.

ORCID

Aaron R. Caldwell http://orcid.org/0000-0002-4541-6283

References

[1] Nalimov VV. In the labyrinths of language: a mathematician's journey. Philadelphia, PA: iSi Press; 1981.

[2] Weissgerber TL, Garovic VD, Milin-Lazovic JS, et al. Reinventing biostatistics education for basic scientists. PLOS Biol. 2016;14:e1002430.

[3] Gigerenzer G. Statistical rituals: the replication delusion and how we got there. Adv Methods Pract Psychol Sci. 2018;1:198–218.

[4] de Groot AD. The meaning of "significance" for different types of research. [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]. 1969. Acta Psychol (Amst). 2014;148: 188–194.

[5] Nickerson RS. Confirmation bias: A ubiquitous phenomenon in many guises. Rev Gen Psychol. [Internet]. 1998 [cited 2018 Aug 3];2:175–220. Available from: http://0-search.ebscohost.com.library.uark.edu/login.aspx?direct=true&db=pdh&AN=1998-02489-003&site=ehost-live&scope=site

[6] Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci. 2011;22:1359–1366.

[7] Heininga VE, Oldehinkel AJ, Veenstra R, et al. I just ran a thousand analyses: benefits of multiple testing in understanding equivocal evidence on gene-environment interactions. PloS One. 2015;10:e0125383.

[8] Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol. 2015;68:1046–1058.

[9] Nosek BA, Ebersole CR, DeHaven A, et al. The preregistration revolution. Proc Natl Acad Sci U S A. [Internet]. 2017 [cited 2018 Jun 19]. Available from: http://www.pnas.org/content/early/2018/03/08/1708274114#ref-3

[10] Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: part I. Biometrika. 1928;20A:175–240.

[11] Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: part II. Biometrika. 1928;20A:263–294.

[12] Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond Ser Contain Pap Math Phys Charact. 1933;231:289–337.

[13] Salsburg D. The lady tasting tea: how statistics revolutionized science in the twentieth century. New York, NY: Henry Holt and Company; 2001.

[14] Fisher RA. The statistical method in psychical research. Proceedings of the Society for Psychical Research. 1929;39:388–391.

[15] Lakens D, Adolfi FG, Albers CJ, et al. Justify your alpha. Nat Hum Behav. 2018;2:168–171.

[16] Mayo DG. Statistical inference as severe testing: how to get beyond the statistics wars [Internet]. New York, NY: Cambridge University Press; 2018 [cited 2018 Oct 11]. DOI:10.1017/9781107286184

[17] Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods. 2000;5:241–301.

[18] Pernet C. Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice. F1000Res. 2017;4:621.

[19] Curran-Everett D. Explorations in statistics: hypothesis tests and P values. Adv Physiol Educ. 2009;33:81–86.

[20] Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. PeerJ. 2017;5:e3544.

[21] Cohen J. The earth is round ($p < .05$): rejoinder. Am Psychol. 1995;50:1103.

[22] Kruschke J. Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan. Waltham, MA: Academic Press; 2014.

[23] Gelman A, Carlin JB, Stern HS, et al. Bayesian data analysis. Third ed. Boca Raton, FL: CRC Press; 2013.

[24] Mengersen KL, Drovandi CC, Robert CP, et al. Bayesian estimation of small effects in exercise and sports science. Plos One. 2016;11:e0147311.

[25] Royall R. Statistical evidence: a likelihood paradigm. Boca Raton, FL: CRC Press; 2017.

[26] Aitkin M. Statistical modelling: the likelihood approach. J R Stat Soc Ser Stat. 1986;35:103–113.

[27] Dienes Z. Understanding psychology as a science: an introduction to scientific and statistical inference. London, UK: Palgrave-Macmillan; 2008.

[28] Cumming G. The new statistics: why and how. Psychol Sci. 2014;25:7–29.
[29] McGuire WJ. A perspectivist approach to theory construction. Personal Soc Psychol Rev Off J Soc Personal Soc Psychol Inc. 2004;8:173–182.
[30] Popper K. Realism and the aim of science: from the post-script to the logic of scientific discovery. Abingdon-on-Thames, UK: Routledge; 2013.
[31] Lakatos I. The methodology of scientific research programmes: volume 1: philosophical papers. Cambridge, UK: Cambridge University Press; 1978.
[32] Meehl PE. Theory-testing in psychology and physics: a methodological paradox. Philos Sci. 1967;34:103–115.
[33] Lloyd A, Havenith G. Interactions in human performance: an individual and combined stressors approach. Temp Multidiscip Biomed J. 2016;3:514–517.
[34] Mayo DG, Spanos A. Severe testing as a basic concept in a Neyman–pearson philosophy of induction. Br J Philos Sci. 2006;57:323–357.
[35] Greenhaff PL. Cardiovascular fitness and thermoregulation during prolonged exercise in man. Br J Sports Med. 1989;23:109–114.
[36] Jay O, Bain AR, Deren TM, et al. Large differences in peak oxygen uptake do not independently alter changes in core temperature and sweating during exercise. Am J Physiol Regul Integr Comp Physiol. 2011;301:R832–841.
[37] Lakatos I. Criticism and the methodology of scientific research programmes. Proc Aristot Soc. 1968;69:149–186.
[38] Jay O, Cramer MN. A new approach for comparing thermoregulatory responses of subjects with different body sizes. Temp Austin Tex. 2015;2:42–43.
[39] Cheuvront SN. Match maker: how to compare thermoregulatory responses in groups of different body mass and surface area. J Appl Physiol Bethesda Md 1985. 2014;116:1121–1122.
[40] Abdi H. Coefficient of variation. Encycl Res Des. 2010;1:169–171.
[41] Cheuvront SN, Ely BR, Kenefick RW, et al. Biological variation and diagnostic accuracy of dehydration assessment markers. Am J Clin Nutr. 2010;92:565–573.
[42] Cheuvront SN, Kenefick RW. CORP: improving the status quo for measuring whole body sweat losses. J Appl Physiol Bethesda Md 1985. 2017;123:632–636.
[43] Stein RJ, Haddock CK, Poston WSC, et al. Precision in weighing: a comparison of scales found in physician offices, fitness centers, and weight loss centers. Public Health Rep Wash DC 1974. 2005;120:266–270.
[44] Travers GJS, Nichols DS, Farooq A, et al. Validation of an ingestible temperature data logging and telemetry system during exercise in the heat. Temp Austin Tex. 2016;3:208–219.
[45] Silberzahn R, Uhlmann EL, Martin DP, et al. Many analysts, one data set: making transparent how variations in analytic choices affect results. Adv Methods Pract Psychol Sci. 2018;1(3). 2515245917747646.
[46] Coll M-P. Meta-analysis of ERP investigations of pain empathy underlines methodological issues in ERP research. Soc Cogn Affect Neurosci. 2018;13(10): nsy072.
[47] Gelman A, Loken E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. [Internet]. 2013. [Cited July 14, 2018]. Available from: https://stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
[48] Sainani K. Interpreting "Null" results. Phys Med Rehabil. 2013;5:520–523.
[49] Campbell H, Gustafson P. Conditional equivalence testing: an alternative remedy for publication bias. PLoS ONE. [Internet]. 2018 [cited 2018 Aug 14]; 13. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5898747/
[50] Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc Psychol Personal Sci. 2017;8:355–362.
[51] Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: a tutorial. Adv Methods Pract Psychol Sci. 2018;1:259–269.
[52] Lakens D, McLatchie N, Isager PM, et al. Improving inferences about null effects with Bayes factors and equivalence tests. J Gerontol B Psychol Sci Soc Sci. 2018.
[53] Wellek S. Testing statistical hypotheses of equivalence and noninferiority. [Internet]. Boca Raton, FL: CRC Press; 2010. Available from: https://www.crcpress.com/Testing-Statistical-Hypotheses-of-Equivalence-and-Noninferiority/Wellek/p/book/9781439808184
[54] Lakens D. TOSTER: Two One-Sided Tests (TOST) equivalence testing. [Internet]. 2018 [cited 2018 Aug 29]. Available from: https://CRAN.R-project.org/package=TOSTER.
[55] Wellek S, Ziegler P. EQUIVNONINF: testing for equivalence and noninferiority [Internet]. 2017 [cited 2018 Aug 29]. Available from: https://CRAN.R-project.org/package=EQUIVNONINF.
[56] Greenland S, Poole C. Living with p values: resurrecting a Bayesian perspective on frequentist statistics. Epidemiol Camb Mass. 2013;24:62–68.
[57] Sainani KL. The problem with "magnitude-based inference". Med Sci Sports Exercise. [Internet]. 2018. Available from: http://europepmc.org/abstract/med/29683920
[58] Mansournia MA, Altman DG. Some methodological issues in the design and analysis of cluster randomised trials. Br J Sports Med. 2019;53(9):573–557. sbjsports-2018–099628.
[59] Lakens D, Delacre M. Equivalence testing and the second generation P-value. PsyArXiv [Internet]. 2018 [cited 2018 Aug 29]; Available from: https://psyarxiv.com/7k6ay/.

[60] Curran-Everett D. Explorations in statistics: the assumption of normality. Adv Physiol Educ. 2017;41:449–453.
[61] Curran-Everett D. Explorations in statistics: the log transformation. Adv Physiol Educ. 2018;42:343–347.
[62] Curran-Everett D. Explorations in statistics: permutation methods. Adv Physiol Educ. 2012;36:181–187.
[63] Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979;7:1–26.
[64] Lund A, Lund M. Statistical Test Selector | Laerd Statistics Premium [Internet]. [cited 2018 Aug 29]. Available from: https://statistics.laerd.com/premium/sts/index.php.
[65] Rosner B. Fundamentals of biostatistics [Internet]. 8th ed. Cengage Learning: US; 2015 [cited 2018 Aug 30]. Available from: https://cengage.com.au/product/title/fundamentals-of-biostatistics/isbn/9781305268920
[66] Curran-Everett D. CORP: minimizing the chances of false positives and false negatives. J Appl Physiol Bethesda Md 1985. 2017;122:91–95.
[67] Curran-Everett D. Explorations in statistics: confidence intervals. Adv Physiol Educ. 2009;33:87–90.
[68] Curran-Everett D. Explorations in statistics: the bootstrap. Adv Physiol Educ. 2009;33:286–292.
[69] Curran-Everett D. Explorations in statistics: the analysis of ratios and normalized data. Adv Physiol Educ. 2013;37:213–219.
[70] Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14:365–376.
[71] Havenith G. Individualized model of human thermoregulation for the simulation of heat stress response. J Appl Physiol. 2001;90:1943–1954.
[72] Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19:231–240.
[73] Cook JA, Hislop J, Adewuyi TE, et al. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. NIHR J Lib. 2014;18.
[74] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10:407–415.
[75] Prentice DA, Miller DT. When small effects are impressive. Psychol Bull. 1992;112:160–164.
[76] Albers C, Lakens D. When power analyses based on pilot data are biased: inaccurate effect size estimators and follow-up bias. J Exp Soc Psychol. 2018;74:187–195.
[77] Anderson SF, Kelley K, Maxwell SE. Sample-size planning for more accurate statistical power: a method adjusting sample effect sizes for publication bias and uncertainty. Psychol Sci. 2017;28:1547–1562.
[78] Faul F, Erdfelder E, Lang A-G, et al. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39:175–191.
[79] Westfall J. PANGEA: Power ANalysis for GEneral Anova designs [Internet]. [Cited 2018 Sep 18]. 2016. Available from: https://jakewestfall.shinyapps.io/pangea/.
[80] Lakens D, Caldwell A. ANOVA simulation [Internet]. [Cited 2018 Oct 18]. 2018. Available from: http://shiny.ieis.tue.nl/anova_power/.
[81] Champely S, Ekstrom C, Dalgaard P, et al. Pwr: basic functions for power analysis. [Internet]. 2018 [cited 2018 Aug 3]. Available from: https://CRAN.R-project.org/package=pwr.
[82] HyLown Consulting LLC. Power and sample size | free online calculators. [Internet]. [cited 2018 Aug 6]. Available from: http://powerandsamplesize.com/.
[83] Kane SP. Sample size calculator. [Internet]. [cited 2018 Aug 6]. Available from: http://clincalc.com/stats/samplesize.aspx.
[84] Blair G, Cooper J, Coppock A, et al. Declare design: declare and diagnose research designs [Internet]. 2018 [cited 2018 Sep 12]. Available from: https://CRAN.R-project.org/package=DeclareDesign.
[85] Anderson SF, Kelley K. BUCSS: Bias and Uncertainty Corrected Sample Size [Internet]. 2018 [cited 2018 Sep 12]. Available from: https://CRAN.R-project.org/package=BUCSS.
[86] PASS 16 power analysis and sample size software [Internet]. Kaysville, Utah, USA: NCSS, LLC.; 2018. Available from: https://www.ncss.com/software/pass/
[87] Kelley K, Maxwell SE. Sample size for multiple regression: obtaining regression coefficients that are accurate, not simply significant. Psychol Methods. 2003;8:305–321.
[88] Kelley K. Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. Behav Res Methods. 2007;39:755–766.
[89] Rothman KJ, Greenland S. Planning study size based on precision rather than power. Epidemiology. 2018;29:599.
[90] Swaen GG, Teggeler O, van Amelsvoort LG. False positive outcomes and design characteristics in occupational cancer epidemiology studies. Int J Epidemiol. 2001;30:948–954.
[91] Lewis M. the undoing project: a friendship that changed our minds. New York, NY: W. W. Norton & Company; 2016.
[92] Kaplan RM, Irvin VL, Garattini S. Likelihood of null effects of large NHLBI clinical trials has increased over time. PLoS ONE. [Internet]. 2015 [cited 2018 Aug 12];10:e0132382. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4526697/
[93] Franco A, Malhotra N, Simonovits G. Social science. Publication bias in the social sciences: unlocking the file drawer. Science. 2014;345:1502–1505.
[94] Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. J R Stat Soc. 1969;132:235–244.
[95] John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol Sci. 2012;23:524–532.

[96] Weber F, Hoang Do JP, Chung S, et al. Regulation of REM and Non-REM sleep by periaqueductal GABAergic neurons. Nat Commun. [Internet]. 2018 [cited 2018 Sep 13];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5783937/
[97] Lakens D. Performing high-powered studies efficiently with sequential analyses. Eur J Soc Psychol. 2014;44:701–710.
[98] Wald A. Sequential analysis. Oxford, England: John Wiley; 1947.
[99] Viele K, McGlothlin A, Broglio K. Interpretation of clinical trials that stopped early. Jama. 2016;315:1646–1647.
[100] Jennison C, Turnbull BW, Turnbull BW. Group sequential methods with applications to clinical trials. [Internet]. Chapman and Hall/CRC; 1999 [cited 2018 Sep 13]. Available from: https://www.taylorfrancis.com/books/9781584888581
[101] Reboussin DM, DeMets DL, Kim KM, et al. Computations for group sequential boundaries using the Lan-DeMets spending function method. Control Clin Trials. 2000;21:190–207.
[102] Pocock SJ. When (Not) to stop a clinical trial for benefit. Jama. 2005;294:2228–2230.
[103] Lakens D, Evers ERK. Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies. Perspect Psychol Sci. 2014;9:278–292.
[104] Maxwell SE, Kelley K, Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. Annu Rev Psychol. 2008;59:537–563.
[105] Kelley K, Rausch JR. Sample size planning for the standardized mean difference: accuracy in parameter estimation via narrow confidence intervals. Psychol Methods. 2006;11:363–385.
[106] Borg DN, Osborne JO, Stewart IB, et al. The reproducibility of 10 and 20km time trial cycling performance in recreational cyclists, runners and team sport athletes. J Sci Med Sport. 2018;21:858–863.
[107] Schönbrodt FD, Wagenmakers E-J, Zehetleitner M, et al. Sequential hypothesis testing with Bayes factors: efficiently testing mean differences. Psychol Methods. 2017;22:322–339.
[108] Ghosh BK, Sen PK. Handbook of sequential analysis. Boca Raton, FL: CRC Press; 1991.
[109] Kelley K. MBESS: the MBESS R Package. [Internet]. 2018 [cited 2018 Aug 3]. Available from: https://CRAN.R-project.org/package=MBESS
[110] Pahl R. GroupSeq: A GUI-based program to compute probabilities regarding group sequential designs [Internet]. 2018 [cited 2018 Sep 13]. Available from: https://CRAN.R-project.org/package=GroupSeq.
[111] Sainani KL. A checklist for analyzing data. Pm&R. 2018;10:963–965.
[112] Field A, Miles J, Field Z. Discovering statistics using R. Thousand Oaks, CL: SAGE Publications; 2012.
[113] Wilcox RR. Introduction to Robust estimation and hypothesis testing. 3rd ed. Waltham: Academic Press; 2012.
[114] Good PI. Permutation, parametric, and bootstrap tests of hypotheses. 3rd ed. New York, NY: Springer; 2004.
[115] Grubbs FE. Procedures for detecting outlying observations in samples. Technometrics. 1969;11:1–21.
[116] Tietjen GL, Moore RH. Some Grubbs-type statistics for the detection of several outliers. Technometrics. 1972;14:583–597.
[117] Rosner B. Percentage points for a generalized ESD many-outlier procedure. Technometrics. 1983;25:165–172.
[118] Wilcox RR. Fundamentals of modern statistical methods: substantially improving power and accuracy. New York, NY: Springer-Verlag; 2001.
[119] Lindsey ML, Gray GA, Wood SK, et al. Statistical considerations in reporting cardiovascular research. Am J Physiol Heart Circ Physiol. 2018;315:H303–H313.
[120] Sainani KL. Dealing with missing data. Pm&R. 2015;7:990–994.
[121] Marmolejo-Ramos F, Cousineau D, Benites L, et al. On the efficacy of procedures to normalize Ex-Gaussian distributions. Front Psychol. [Internet]. 2015 [cited 2018 Sep 28];5. Available from : https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01548/full
[122] Mangiafico S. Rcompanion: functions to support extension education program evaluation [Internet]. 2018 [cited 2018 Oct 4]. Available from: https://CRAN.R-project.org/package=rcompanion.
[123] Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Ser B Methodol. 1964;26:211–252.
[124] Rousselet GA, Wilcox RR. Reaction times and other skewed distributions: problems with the mean and the median. bioRxiv. 2018;383935.
[125] Kirby KN, Gerlanc D. BootES: an R package for bootstrap confidence intervals on effect sizes. Behav Res Methods. 2013;45:905–927.
[126] Mair P, Wilcox R. WRS2: A collection of robust statistical methods [Internet]. 2018 [cited 2018 Sep 24]. Available from: https://CRAN.R-project.org/package=WRS2
[127] Garren ST. jmuOutlier: permutation tests for nonparametric statistics. [Internet]. 2018 [cited 2018 Sep 24]. Available from: https://CRAN.R-project.org/package=jmuOutlier.
[128] Erceg-Hurn DM, Mirosevich VM. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. Am Psychol. 2008;63:591–601.
[129] Vieth E. Fitting piecewise linear regression functions to biological responses. J Appl Physiol. 1989;67:390–396.
[130] Pedhazur E. Multiple regression in behavioral research: explanation and prediction. 3rd ed. United States: Thomason Learning; 1997.
[131] Richmond VL, Davey S, Griggs K, et al. Prediction of core body temperature from multiple variables. Ann Occup Hyg. 2015;59:1168–1178.

[132] Vigotsky AD, Schoenfeld BJ, Than C, et al. Methods matter: the relationship between strength and hypertrophy depends on methods of measurement and analysis. PeerJ. 2018;6:e5071.
[133] Curran-Everett D. Multiple comparisons: philosophies and illustrations. Am J Physiol Regul Integr Comp Physiol. 2000;279:R1–R8.
[134] Kelley K, Preacher KJ. On effect size. Psychol Methods. 2012;17:137–152.
[135] Gagge AP, Stolwijk JAJ, Hardy JD. Comfort and thermal sensations and associated physiological responses at various ambient temperatures. Environ Res. 1967;1:1–20.
[136] Thomas JR, Salazar W, Landers DM. What is missing in p less than .05? Effect size. Res Q Exerc Sport. 1991;62:344–348.
[137] Thomas JR, Lochbaum MR, Landers DM, et al. Planning significant and meaningful research in exercise science: estimating sample size. Res Q Exerc Sport. 1997;68:33–43.
[138] Rhea MR. Determining the magnitude of treatment effects in strength training research through the use of the effect size. J Strength Cond Res. 2004;18:918–920.
[139] Quintana DS. Statistical considerations for reporting and planning heart rate variability case-control studies. Psychophysiology. 2017;54:344–349.
[140] Schönbrodt FD, Perugini M. At what sample size do correlations stabilize? J Res Pers. 2013;47:609–612.
[141] Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. J Exp Psychol Gen. 2012;141:2–18.
[142] Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc. 2007;82:591–605.
[143] Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. [Internet]. 2017 [cited 2019 Feb 24]. Available from: https://CRAN.R-project.org/package=boot.
[144] Cumming G. ESCI (Exploratory Software for Confidence Intervals). [Internet]. Introd. New Stat. 2016 [cited 2018 Sep 24]. Available from: https://thenewstatistics.com/itns/esci/.
[145] Del Re A. Compute.es: compute effect sizes. [Internet]. 2014 [cited 2018 Sep 24]. Available from: https://CRAN.R-project.org/package=compute.es.
[146] Cumming G, Finch S. A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. Educ Psychol Meas. 2001;61:532–574.
[147] Hoekstra R, Morey RD, Rouder JN, et al. Robust misinterpretation of confidence intervals. Psychon Bull Rev. 2014;21:1157–1164.
[148] International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. Ann Int Med. 1997;126:36–47.
[149] Abelson RP. A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). In: Lisa L. Harlow, editor. What there were no significance tests. Mahwah, NJ: Erlbaum; 1997. p. 472.
[150] Curran-Everett D, Benos DJ. Guidelines for reporting statistics in journals published by the American Physiological Society. Adv Physiol Educ. 2004;28:85–87.
[151] American Psychological Assocation. Publication manual of the American Psychological Association. 6th ed. Washington, DC: American Psychological Association; 2009.
[152] Weissgerber TL, Milic NM, Winham SJ, et al. Beyond bar and line graphs: time for a new data presentation paradigm. PLOS Biol. 2015;13:e1002128.
[153] Nuijten MB, Hartgerink CHJ, van Assen MALM, et al. The prevalence of statistical reporting errors in psychology (1985–2013). Behav Res Methods. 2016;48:1205–1226.
[154] Hardwicke TE, Ioannidis JPA. Populating the Data Ark: an attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. Plos One. 2018;13:e0201856.
[155] Steegen S, Tuerlinckx F, Gelman A, et al. Increasing transparency through a multiverse analysis. Perspect Psychol Sci. 2016;11:702–712.
[156] Tufte ER. The visual display of quantitative information. 2nd ed. Cheshire, CT: Grpahics Press; 2001.
[157] Weissgerber TL, Garovic VD, Savic M, et al. From static to interactive: transforming data visualization to improve transparency. PLOS Biol. 2016;14:e1002484.
[158] Tversky A, Kahneman D. Belief in the law of small numbers. Psychol Bull. 1971;76:105–110.