Appetite xxx (xxxx) xxx

Contents lists available at ScienceDirect

Appetite

journal homepage: www.elsevier.com/locate/appet

Editorial

Guidelines on design, measurement and statistics for *Appetite*

1. Introduction

Appetite strives to publish the highest quality science possible. To that end, we offer the following Guidelines on experimental design, quantitative measurement, and descriptive and analytic statistics.

Authors should pay particular attention to the reporting guidelines. For the reasons described below, *Appetite* encourages detailed reporting of statistics. It is often convenient to provide this information as supplementary files (*Appetite*'s term for appendices).

The Guidelines describe our understanding of current best practices for the commonest methods in appetite research. In most cases we provide background to clarify the provenance of the practices. We understand that best practices are often aspirational goals and that there are reasonable rationales for not always fulfilling them.

At the same time, false-positive publications are a serious problem in psychology (e.g.,Brown et al., 2018; Ioannidis, 2005, 2014; Landis et al., 2012; Nelson et al., 2018; Simmons et al., 2011; Wasserstein et al., 2019). False-positive results are often related to poor statistical practice. *Appetite* seeks to minimize such issues. To this end, authors are urged to: [i] design the experiment, including the statistical approach, in advance; [ii] conduct the research – including the statistics – with integrity; [iii] fully and clearly describe the design and execution of the experiments, including statistical methods, randomized or blinded aspects of the design, loss of data, etc.; and [iv] interpret statistical outcomes in an enlightened fashion, as discussed in §4.2.

Because the Guidelines are not comprehensive, authors are advised to consult statisticians for further guidance concerning the design and analysis of their own work. In addition, the development of many of the statistical approaches described here are themselves active research areas, which is another good reason to consult professionals. The *BMJ*'s "Statistics Notes" series (www.bmj.com/specialties/statistics-notes), which has run since 1994, and the *American Journal of Clinical Nutrition*'s series "Best (but Oft-Forgotten) Practices" (beginning with Bier et al., 2015) are helpful resources. We also recommend that authors keep abreast of the literature on how best to mitigate the problem of false-positive results (e.g., Ioannidis, 2014; Nelson et al., 2018; Simmons et al., 2011; Wasserstein et al., 2019).

2. Experimental design and related topics

2.1. Prespecification

Experimental designs and statistical approaches should be specified in advance. For a comprehensive discussion, see Nosek et al. (2018). Preregistration services include the Center for Open Science (https://osf.io/prereg/) and ClinicalTrials.gov (https://www.clinicaltrials.gov/). Other alternatives listed by region are at www.who.int/ictrp/network/primary/en/.

2.2. Ethics

Research involving human participants, human material, or human data, must have been performed in accordance with the Declaration of Helsinki and must have been reviewed by an appropriate independent ethics committee. Similarly, research involving non-human animals or material derived from them must have been approved by an appropriate independent ethics committee. A statement detailing the approval, including the name of the ethics committee and the reference number where appropriate, must appear in the manuscript.

2.3. Planning for meta-analyses

Scientific meaning is rarely established by a single study, but rather by the cumulative effect of many similar studies. The state-of-the-art for the quantitative integration of similar studies is meta-analysis (Borenstein et al., 2009; Cooper, 2010). Thus, a useful criterion for full data reporting is for authors to plan for the potential later inclusion of their work in meta-analyses, i.e., quantitative integration with other similar studies. To meet this criterion, all sample sizes, measures and outcome estimates (means, etc.), and their variabilities should be reported. If this does not fit easily with the chosen style of presentation, it should be included as supplementary files.

2.4. Conducting meta-analyses

Authors performing meta-analyses should adhere to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; www.prisma-statement.org).

2.5. Studies with humans

Reports of clinical trials should adhere to the Consolidated Standards of Reporting Trials (CONSORT; www.consort-statement.org) and applicable updates; for example, updates concerning studies of cluster randomized trials (Campbell et al., 2012), of non-pharmacological treatments, which includes guidance on reducing bias when blinding is not possible (Bourtron et al., 2017), and of randomized crossover designs (Dwan et al., 2019). Much of the guidance contained in the CONSORT statement also applies to the smaller experimental studies with human participants. Guidance available at https://www.equat

https://doi.org/10.1016/j.appet.2021.105731

Available online 2 October 2021

or-network.org/ includes other types of studies with humans.

2.6. Studies with non-human animals

Reports of studies with non-human animals should adhere to the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines (https://arriveguidelines.org/) (Percie du Sert et al., 2020).

2.7. Robust statistical methods

Computing has permitted the development of a variety of novel and powerful statistical methods, commonly known as robust statistics (Wilcox, 2003). Two useful methods are computerized resampling and bootstrapping (Kirby & Gerlanc, 2013). Authors are encouraged to consider these alternatives.

2.8. Extreme values/Outliers

Robust statistical methods to detect and exclude extreme values or outliers are often useful in small-sample studies. Robust methods minimize the influence of suspected outliers on the statistic used to identify them. A simple robust method is to compute the probability of suspected extreme values using median standard scores:

(x – group median)/1.48 MAD

where x is the suspect datum, MAD is the median absolute deviate (median of $|x_i$ – group median| for each x_i in the group; note: 1.48 MAD ≈ the group's standard deviation [SD]). Leys et al. (2013) provide instructions for computing MAD in the SPSS and R statistical packages. Other robust methods are described by Rousseeuw and Croux (1993).

2.9. Null results

Appetite recognizes the need to publish well designed experiments that address interesting questions but fail to result in convincing outcomes. Not to do so inflates the meaning of positive reports and invalidates future meta-analyses. Negative data are rarely considered suitable for publication, however, if the experimental design does not include a suitable power analysis.

It is crucial to understand that "negative data" does not mean that the statistics show that there is no difference. Rather, it means only that the statistics failed to demonstrate evidence of a difference, which is very different. As has been pointed out repeatedly, "absence of evidence is not evidence of absence" (Alderson, 2004; Altman & Bland, 1995; Bramness et al., 2008; Hartung et al., 1985) Negative data should be described with this in mind. Bayesian statistics (§4.1.) are especially useful for negative data as they enable quantification of the strength of negative data.

2.10. Descriptive, exploratory and analytic statistics

Descriptive statistics summarize the data, and analytic statistics (§5–7) assist in making inferences about the meaning of data. Between the two lies exploratory data analysis or data mining, which refers to attempts to understand the collected data using a variety of descriptive approaches with the goal of discovering unexpected possibilities that could guide future experiments (Gelman, 2003, Tufte, 2001; Tukey, 1977; Wainer, 2007). Wainer and Velleman's (2008) exploration of blood glucose level graphing is an excellent example. Recently, nonparametric estimation methods have been used to quantify exploratory data analysis in novel ways (Harpole et al., 2014). Serendipity plays an important role in science. Exploratory analyses are welcome, but should be clearly labeled as such and described separately from analytic statistics.

2.11. Data deposition

Both raw and analyzed data should be maintained and made available upon request. Stored data should be organized and coded so that it is comprehensible. Authors are encouraged to deposit data in public repositories; a registry of public repositories is: www.re3data.org/

2.12. Reporting

Authors should clearly describe the design and execution of the experiments, including all measures, data manipulations, and data exclusions. All randomized or blinded aspects of the design should be mentioned. Authors should clearly state whether the analytic plan was prespecified and describe any deviations from it. Studies in which the analytic strategy is not prespecified should be labeled as exploratory.

3. Measurement

3.1. Introduction

In the physical sciences, "fundamental measurements" are generated when quantities that can be ordered are found to correspond to a number of units of an unvarying physical standard. For example, people's heights can be measured with meter sticks placed end to end, and their weights can be measured with a balance pan to which a number of standard weights can be added. Such measurement operations are called concatenations (Bond & Fox, 2015; Campbell, 1920; Tal, 2020). Derived measures, such as body mass index (weight in kg/height in m^2), are computed from two or more fundamental measures. Fundamental and derived measures can be expressed as distances along a line of continuous, infinitely divisible numbers with evenly spaced multiples of any number and a meaningful 0; i.e., they are real numbers, designated $\mathbb{R}$ in math.

In psychology and many other fields there are no such measurements. Although quantifiable, orderable constructs abound, there are no physical standards for their measurement. This leads to issues around the concepts of measurement error and measurement scales that authors must cope with.

3.2. Measurement error

Measurement error refers to the resolution of the measurement standard. The lack of physical measurement standards in psychology is conducive to ignoring measurement error. This should not be done.

Measurement error is usually expressed as the SD of the theoretical distribution of measures. The SD of a uniform distribution from a to b = $(b - a)/12^{0.5}$. For measurements in the form 1, 2, 3, …, if integers correspond to values of a latent variable that is located along a segment of a continuous real-number scale, then 1) the absolute error is ±0.5 units of the integer, and 2) b – a = 1, so the SD of measurement error = 0.29.

Measurement error propagates through mathematical manipulations according to Gauss' theory of errors (Joint Committee for Guides in Metrology, 2008). If one adds measurements x_1, x_2, …x_n that have measurement errors SD_1, SD_2, …SD_n, then the measurement error of the sum = $(SD_1^2 + SD_2^2, \ldots + SD_n^2)^{0.5}$. If one multiplies or divides a measure by a constant, the SD of measurement error is simply multiplied or divided by the same constant. Thus, the measurement error of a mean is less than that of the individual measurements.

3.3. Measurement scales

To address the lack of fundamental measures in psychology and many other fields, Stevens (1946) introduced an expanded categorization of measurement scales and described the mathematical operations appropriate to each. The definitions are given in Table 1.

According to these definitions, fundamental and derived measures as defined in §3.1. are ratio scales; i.e., real numbers. Interval scales are segments of the real number line. Ordinal scales, and improvements to them (Coombs, 1950; Guttman, 1944; Kyngdon, 2013), are not real numbers. Nominal (or categorical) measures are not numbers in any mathematical sense, but merely names of groups. They should perhaps be renamed nominal properties (Adroher et al., 2018).

Stevens (1946) accepted that few of psychology's measures were interval or ratio scale (some psychophysical measures were apparent exceptions). Ironically, however, he also initiated the long tradition of using all the tools of arithmetic to analyze presumably ordinal data anyway. He defended this pragmatically, arguing that the outcomes were often useful. Others pointed out that studies of convergent and construct validity suggested that the errors produced by treating ordinal data as interval are usually relatively small. These perspectives remain the common practice of our field.

Within ~20 y of Stevens (1946) paper, a new measurement theory, called conjoint measurement, was developed that provided mathematically valid alternatives to Stevens' pragmatic approach to the problem of ordinal-level measurement (Krantz et al., 1971; Luce and Tukey, 1964; Rasch, 1960). One form of conjoint measurement, item response theory (IRT), is based on the structure of ordered matrices of individual performance on each scale item. The most widely accepted IRT model is Rasch analysis (Adroher et al., 2018; Bond and Fox, 2015). In the simplest case, in which items probing some latent variable are answered or performed successfully or not, Rasch analysis is based on the probability function:

$$P_{nj}(\text{success}) = f(B_n - D_j)$$

where P_{nj}(success) is the probability that participant n responds successfully to item j, B_n is the participant n's ability with respect to the latent variable measured, and D_j is the difficulty of item j. Rasch (1960) demonstrated that if *f* is a logistic function, then the scale has interval properties. Rasch and others have provided similar demonstrations for a number of designs (for further explanation and examples, see Bond and Fox, 2015; da Rocha et al., 2013; Pallant and Tennant, 2007; Sijtsma, 2011). The R statistical package includes extensive software for Rasch modelling (Mair et al., 2017; R Core Team, 2021).

The Rasch model provides statistics quantifying the fits of individual items to the logistic model; i.e., to an interval-level scale. Ideally, scale construction is done iteratively, as adding or omitting items affects the logistic fit of the remaining items. Rasch analysis of existing scales usually indicates that some adjustment improves the fit of the scale to the model.

Conjoint measurement methodology has not penetrated noticeably into appetite research. An exception is the corrected Eating Disorder Examination Questionnaire (Gideon et al., 2016; Jinbo et al., 2021; Kampen, 2019; Prnjak et al., 2020), which should be adopted. There are understandable reasons for the rarity of conjoint measurement in appetite research. Rasch and similar methods require quite large numbers of participants, and appetite research is largely small-scale discovery research, often involving *ad hoc* questionnaire development. Nevertheless, *Appetite* encourages the development of corrected scales, especially for widely used instruments.

Table 1
Stevens' (1946) definitions of scales of measurement.

Scale	Measures derive from determination of:	Permissible arithmetic[a]
Nominal	Equality	None[b]
Ordinal	Greater than or less than	None[c]
Interval	Equality of intervals (e.g., [5–3] = [4–2])	Add, subtract[d]
Ratio	Equality of ratios (e.g., [6/3] = [4/2])	Multiply, divide[e]

[a] Arithmetic operations involving ≥2 measured quantities are permissible if they do not change the properties of the scale; permissible transformations of single measured quantities are described by Sarle (1997).

[b] Counting category members is permitted.

[c] Counting numbers of data points more or less than another is permitted.

[d] Addition and subtraction are permitted. Multiplication and division are permitted on differences relative to a point on the scale, such as the mean, but not on raw data.

[e] Equality of ratios requires that the scale has a meaningful 0 value. Non-linear transformations of data (e.g., squares, roots, logs, etc.) are not permitted.

3.4. Reporting

Data ordinarily should be reported in the form measured, using SI units (Le Système International d'Unités) where possible and clearly defined units otherwise. Data shown in figures or tables should not be described in the text. If the data are in not in natural units (i.e., not g, J, etc.) or not in units with known biological or clinical meaning, then effect sizes are accepted indices of meaningfulness (see §4.2.). The R statistical package calculates effects sizes for a variety of parametric and non-parametric tests (R Core Team, 2021).

Data should be reported using significant figures; i.e., powers of 10 that reflect the precision of the measurements, as discussed in §3.2. The rule of thumb is that measures should be reported to the same precision as the individual measurement (although as described in §3.3, means have greater precision than the sum).

Measurement scales (3.3) determine appropriate forms of data analysis. To describe central tendency, means are appropriate for ratio or interval scales, medians for ordinal scales, and modes for nominal scales. To describe spread, SD and related measures are appropriate for ratio or interval scales, and the index of dispersion (D) is appropriate for ordinal or nominal scales:

$$D = k(n^2 - \text{sum } f_i^2) / n^2 (k - 1)$$

where k is the number of categories or intervals, n is the number of data points, and f is the number of data points in each of the categories, i = 1 to k. Many texts recommend ranges for ordinal data, but this is incorrect because ranges indicate intervals, which are not meaningful for ordinal-scale measures.

In addition to measurement issues, the scientific utility of precision should be considered in reporting. High precision may not be meaningful even if it is accurate. For example, the age of adult participants should not be reported to 0.01 y, which is 3.5 d and has no scientific meaning. Similarly, adult weights reported to precision 0.01 or 0.001 kg are generally meaningless.

4. Introduction to analytic statistics

4.1. Approaches

Statisticians recognize three approaches to analytic statistics. We discuss the two most commonly used approaches, the **statistical-significance approach** and the **estimation approach**. Both are based on the classical contributions to mathematical probability of Carl Friedrich Gauss (distributions of normal errors, least squares estimation, etc.) and Pierre-Simon LaPlace (central limit theorem, etc.) in the early 19th C. Despite their common roots, these involve different analytic methods, different language, and different logical rules for data interpretation. For example, in the statistical significance approach, one tests hypotheses such as, "is there a difference between groups X and Y?", whereas in the estimation approach, rather than framing specific hypotheses, one asks, "how large is the difference between groups X and Y?" *Appetite* accepts both approaches. They should not, however, be mixed in a single experiment.

A third approach is **Bayesian statistics**. In Bayesian statistics data are used to revise a "prior probability" that the population distribution from which the sample is drawn has certain characteristics (e.g., a certain mean) to produce a "posterior probability" (López Puga et al., 2015ab). Prior probabilities can be based on data, reasoning or

speculation. Bayesian statistics quantify the evidence supporting competing hypotheses and as such can be useful in interpreting non-significant results obtained from null hypotheses testing. More specifically, the ratio of the posterior probabilities of alternative hypotheses, known as the Bayes factor, represents the relative weight of evidence in the data for the competing hypotheses. For further reading on the use of the Bayes factor to assess the degree to which non-significant results support a null hypothesis over a theory see Dienes (2014).

Another categorization of analytic statistics is based on assumptions about the population distribution. **Parametric methods** include assumptions about the specific form of the population distribution, for example, that it is Gaussian (normal). In contrast, **nonparametric methods** require few assumptions about population distribution. Guidelines for these approaches are described in §5–7.

4.2. Probabilities and meaning

Both the statistical-significance approach and the estimation approach lead to estimates of the probabilities related to the observed outcomes. In the former, the null hypothesis of no group difference is rejected, and the observed difference is considered statistically significant, if the observed data indicate that the probability that the null hypothesis is true is less than some pre-selected probability, α, most often $P = 0.05$. In the latter, attention is focused on probabilities that various parameters fall in a certain range, usually the range in which the outcome will occur with $P \geq 0.95$, known as the 95% confidence interval or 95% CI.

Statisticians find that interpretations of statistical significance and estimations outcomes are often faulty. Specifically, interpretations often fail to recognize that the probabilities generated are points on a continuous probability continuum, not categorical criteria for dichotomizing results as meaningful or not. Thus, in a statistical significance approach, $P = 0.051$ is usually not meaningfully different from $P = 0.049$, and in an estimation approach, a value associated with P=0.94, i.e., a value within the 95% CI, is not meaningfully different from a value with P=0.96. This kind of misunderstanding has had ripple effects that adversely affect science. In response, the American Statistical Association published a series of articles discussing strategies to mitigate the misuse of statistical significance (Wasserstein et al., 2019). Some important points were:

- Statistical results should be recognized as being incomplete and uncertain (Amrhein et al., 2019).
- Because P values are estimates, they should be reported as exact estimates (e.g., $P = 0.04$ or $P = 0.07$).
- Interpretations should consider the magnitude of differences with respect to behavioral or physiological importance or, if that is not possible, with respect to effect sizes (Blume et al., 2019; Betensky 2019; Goodman, Spruill, & Komaroff, 2019). (Effect sizes are described in several sections below.)
- A finding of statistical significance is not sufficient evidence to conclude that the effect is highly probable, true, or important. Interpretations should consider the outcomes of similar published studies and should recognize that important results will be refined by meta-analyses (see §2.3–2.4) and by experiments with improved measures, more sensitive designs, and larger samples.
- Authors should recognize that subjectivity can influence every step between planning a study to interpreting the results. Therefore, authors should search for and minimize bias in their reasoning and choices thoughtfully (Ioannidis, 2019).

Appetite endorses these and other best practices in analytical statistics.

5. Parametric analytic statistics

5.1. t-tests and ANOVA

The most familiar analytic statistics, t-tests and analysis of variance (ANOVA), are categorical parametric statistics: *categorical* because the independent variable is different levels of some nominal or categorical measure (e.g., two sexes) rather than a continuous dimension (e.g., age), and parametric because they are based on mathematics assuming the Gaussian (normal) distribution and therefore require interval- or ratio-scale measurements. Collapsing dimensional data into categories to enable ANOVA should be avoided.

ANOVA analyses can be re-cast in part or whole as correlational analyses. For example, analysis of covariance (ANCOVA) combines an ANOVA approach with a correlational approach. If a design includes baseline data, considering the experimental data as correlates of the baseline measure is usually the best strategy (George et al., 2016). More complex designs are now frequently analyzed using only correlational approaches, such as GLM (see §5.3).

5.1.1. Assumptions

For t-tests, computer modeling has demonstrated that the assumption that the data are drawn from Gaussian distributions is not crucial; there is little risk of error as long as the distributions are unimodal and fairly symmetric. This is not the case for ANOVA. Rather, the distributions of all groups should be approximately Gaussian unless sample size is at least moderate (often defined as ≥ 30), their variances should be similar, groups sizes should be nearly equal (this is not crucial for one-way ANOVA), and, for repeated measures designs, the sphericity criterion should be met. ANCOVA has the additional requirement that the continuous variables produce parallel correlations. Many computer statistics packages include tests of these criteria. If the assumptions of parametric categorical approaches are not met, non-parametric approaches are called for (see §7).

5.1.2. Data transformation

If the ANOVA assumption of Gaussian distributions is not met, it is common practice to transform the data into a form that does approximate a Gaussian distribution, for example, by using square roots or logs of the data. It should be recognized that this practice comes at a cost. The transforms are typically not permissible operations on the measures as described in §3.3 because they distort the measurement scale. Thus, if the measurements were assumed to be conjoint measures of a certain latent variable, then the transformed data no longer are. This may complicate interpretation in terms of the latent variable, comparison of the outcomes with other analyses in which transforms were not done, etc. For example, in factorial designs, data transforms often produce interactions even if the raw data are additive (see §5.1.3.).

Ratios of random variables are problematic from several perspectives, so that correlational analyses (§5.2) are usually the better choice (Allison, Paultre, Goran, Poehlman, & Heymsfield, 1995). Transformations into percentages of baseline values are especially troublesome because percent changes of small absolute differences relative to smaller baseline values can be larger than percent changes of larger absolute differences relative to larger baseline values. Therefore, percentages should not be done without a theoretical justification. As described above ANCOVA is usually preferable.

5.1.3. Interaction effects

Factorial ANOVA are almost universally analyzed by partitioning the variance among main effects, interaction effects and error, although it is entirely possible to partition variance without interactions. The choice whether to include interaction effects should be an educated one. First, because interactions are defined as departures from additivity, unless the factors are themselves additive, computing additive interactions makes little sense. Second, if the independent variable is truly

categorical, whether its levels are additive or not is impossible to determine; (Caudle and Williams, 1993; Geary, 2013; Winer, 1971).

5.1.4. ANOVA follow-up

ANOVA and related approaches to analyze experiments involving more than two groups are known as omnibus procedures because they yield overall estimates of statistical significance. These usually require follow-up tests to identify the specific source(s) of significance. Unless the experiment is considered exploratory, follow-up tests should protect the experiment- or analysis-wide α (see §5.1.5.).

5.1.5. Multiplicity

If several measures are used to test a single hypothesis (for example, different measures of the same underlying process), these should be regarded as a single family of tests, and it is necessary to maintain or protect the family-wide type 1 error rate (α, the probability of obtaining statistical significance when in fact there is no effect) or, alternatively, the false-discovery rate. In the absence of a hypothesis, descriptive rather than analytic statistics are usually preferable.

Type-1 error rates increase exponentially with the number of tests of the hypothesis (n). This is easily calculated by subtracting the probability of making no type-1 errors from 1:

$$P[1 \text{ or more type-1 errors}] = 1 - (1 - \alpha)^n$$

For example, if a brain-imaging study tests the hypothesis that a manipulation will increase neural activity in the limbic system, if $\alpha = 0.05$, and if 13 limbic areas are measured, then P[1 or more type-1 errors] > 0.50.

A number of methods have been used to protect the experiment (or analysis)-wide type-1 error rate. Some of these, however, have been determined to be defective and should not be used; these include multiple t-tests, (Fisher's) LSD test, and Dunnett's test. Others are valid, but unnecessarily "conservative," i.e., have poor power. This is the case for both the Tukey HSD test and Bonferroni correction procedure.

For this reason, alternatives to these classic methods are increasingly preferred. These alternatives are based on controlling the false-discovery rate (FDR) rate. (Benjamini, 2010; Benjamini & Hochberg, 1995; Curran-Everett, 2000). Rather than controlling multiplicity by computing the probability of at least one false positive result (i.e., the probability of one type 1 error), FDR methods are based on the estimated ratio of the number of false positive results to the total number of rejections of the null hypothesis, i.e., the sum of correct results and false-positive results. In these methods, effects are compared sequentially to adjusted P levels ranging from α/# tests to α. Hochberg (1988) is the most widely used method.

It is important to note that the Bonferroni procedure and FDR strategies can be applied to parametric and nonparametric analyses alike. It is also important to appreciate the difference between simple and complex follow-up tests: the former are valid only to test individual group means; the latter must be used to test combinations of means, an issue that arises frequently (see §5.1.6.). Note that most computerized statistical packages offer only simple follow-up tests.

5.1.6. Complex follow-up tests

Interaction tests arise in designs comparing in two or more experimental effects. These situations require an explicit test of the difference in the two effects; it does not suffice to show that one effect is significant and the other is not. Complex interactions involve more than two effects, for example comparing two (control – test) effects. In many designs, these are the critical outcomes (Nieuwenhuis et al., 2011). Most computerized statistical packages do not offer such tests. They can be done with the methods mentioned in §5.1.5.

5.1.7. Planned comparisons

Typical ANOVA follow-up tests for differences between pairs (Tukey's HSD test, etc.) often involve a large number of tests (if there are k groups in the ANOVA, there are C(k, 2) = k!/[2 (k-2)!] pairwise contrasts). Protecting the analysis-wide α leads to each comparison having rather low power. If several of these differences are not of interest, planned comparisons provide a more powerful alternative. A simple and adaptable planned-comparison method is to design the necessary comparisons and test them using the Hochberg method (see §5.1.5). In the planned-comparisons approach, ANOVA is used simply to generate an experiment-wide standard error of the difference (SED), not to assess overall significance, according to the formula:

$$SED = [\ 2\ MS_{error} / n]^{1/2},$$

where n is the n per group, not the total n in the analysis.

5.1.8. Power

Power refers to the probability of detecting an effect of a certain size. In the statistical-significance approach, power is defined as 1 - β, where β is the probability of a type 2 error, i.e., not detecting a significant effect when there is one. Experiments should be designed with adequate power. Underpowered experiments reduce the probability both [i] true effects will be detected, and [ii] that significant results reflect true effects (Button et al., 2013). Note also that replicating significant results is expected to require larger sample sizes than used in the original study (Button et al., 2013).

5.1.9. Effect size

The statistical outcome describes the probability that the effect might be observed under the null hypothesis. This does not translate simply into a statement of the magnitude of the effect. Effect-size statistics are designed for that purpose. Most effect sizes are differences normalized by their SD, resulting in dimensionless statistics ranging from 0 to 1. In the case of t-tests, if the two groups have means m_1 and m_2, and have similar sample sizes and variabilities, then the difference between them can be described as Cohen's δ (Cohen, 1988, 1992):

$$\delta = (m_1 - m_2) / SD_{pooled}.$$

Cohen's $\delta \geq 0.2$, ≥0.5, and ≥0.8 are generally considered small, medium, and large effects, respectively. Other effect sizes are applicable to other two-sample cases. Lee (2016) lists several.

For ANOVA, a common effect size is:

$$\eta^2 = SS_{factor} / SS_{total}.$$

$\eta^2 \geq 0.01$, ≥0.06 and ≥0.14 are considered a small, moderate and large effects, respectively (Stevens, 2001).

5.1.10. Reporting

Results of categorical statistical tests should be reported in standard detail; i.e. for ANOVA, report the F value, degrees of freedom(*df*), and probability: F(1,25) = 4.33, P = 0.0x. A precision of 0.01 ordinarily suffices for reporting text statistics. As described in §4.2, exact probabilities should be given rather than P < 0.05. Sample sizes should be given, for example in figure captions. Effect sizes are usually helpful. If tables of statistical outcomes are appropriate, these may be given as supplementary data.

If the Bonferroni or a FDR approach (§5.1.5.) is used, then P values should be corrected to make them comparable to α. For example, if a particular difference is compared against α/3 then the three times the observed P value is comparable to α. These should be referred to as $P_{corrected}$.

Reporting variability brings several choices. Optimally, both the SD as a measure of population spread and either the standard error of the mean (SEM) or 95% CI (assuming α = 0.05) as a measure of the accuracy of the estimation of the mean are reported. Carter (2012) describes the advantages of the 95% confidence interval over the SEM. Note, however, that if data derive from repeated-measures designs, both the usual SEM

and 95% CI conflate within- and between-subject variability; in such cases, the SED (§5.1.7) or repeated-measures CI are more meaningful.

5.2. Correlational analyses

Correlational or dimensional analyses are applicable to a variety of bivariate and multivariate data. Methods for both the statistical-significance approach (described here) and the estimation approach are available. Several correlational methods (mediation analysis, path analysis, etc.) suggest causal relationships, but correlations never prove causality.

5.2.1. Simple regression

For bivariate data (x,y), the regression line minimizes the sum of squared deviations in y from the fit line; x values are considered error-free. The regression line has the form $\hat{y} = b_0 + bx$, where $\hat{y}$ is the predicted y value (the $\hat{y} - y$ values for each x are called residuals), b_0 is the intercept of the regression, and b is its slope. The process generates several parameter estimates, including: 1) an F test of the significance of the regression; 2) 95% CI for each x value; 3) β, which is b in SD units, $\beta = b\ (SD_x/SD_y)$; 4) the effect size or coefficient of determination, R^2, which is the proportion of variance in y explained by variance in x; and 5) the Pearson correlation coefficient, r, which is the square root of R^2; r takes the sign of b, and is another measure of effect size (note that r is not equivalent to δ; rather, $\delta = 2r/(1 - r^2)^{1/2}$). Importantly, none of these parameters gives an impression of what the data actually look like; for that, the scatter plot is indispensable. Anscombe (1973) provided a graphic example of how different data sets with identical b_0, b, variance and r can be.

Different subgroups should not be included in a single correlation unless each group appears to have the same slope and intercept as the overall correlation; failure to do so can lead to "Simpson's paradox" - an overall effect whose direction is opposite to those of the subgroups (Klevit et al., 2013). Collapsing dimensional data into categories to enable categorical analysis approaches (e.g., ANOVA) should be avoided.

5.2.1.1. Assumptions and reporting. The assumptions of simple regression are: 1) the data are interval or ratio scale; 2) the data are independent; 3) the two variables are linearly related; 4) that there are no or few extreme values (if there are outliers, the regression should be verified in their absence); 5) the residuals are normally distributed and homoscedastic. There are no assumptions about the distributions of the independent variable.

Authors should report: 1) how the assumptions were verified; 2) sample size; 3) b_0 and b; 4) signed r, its *df*, and P. The unsigned r or R^2 serves as an effect size.

5.2.2. Multiple regression

Multiple regression describes linear relationships between several independent variables and one dependent variable. A multiple correlation coefficient r is produced, with R^2 indicating the proportion of variance in the dependent variable accounted for by all the independent variables. In multiple regression, the β for each independent variable is a standardized (as in §5.2.1.) partial weight, indicating the unique contribution of that variable after controlling for the effect of all the other variables. Thus, if the dependent variables are correlated with each other, β can be quite small. For this reason, the interpretation of multiple correlations usually requires consideration of both b and β values. Adequate sample size for multiple regression is 5–10 times the number of variables.

Classical mediation analysis is a multiple regression method in which a series of regressions are used to indicate whether a variable intervening between a predictor and an outcome explains part or all of the relationship between the predictor and outcome. If the intervening variable is found to interact with the predictor, then the relationship is referred to as moderation rather than mediation. Classical mediation analysis is increasingly replaced by a bootstrapping method to establish mediation (Hayes, 2017). Bootstrapping is a nonparametric method that has the advantages that normality is not required and that smaller sample sizes can be used.

5.2.2.1. Assumptions and reporting. Multiple regression makes the same assumptions as simple regression (§5.2.1.1.). Again, there are no assumptions about the distributions of the independent variables. If some independent variables are dichotomous, "centering" often renders outcomes more intelligible (Kraemer and Blasey, 2004).

Authors should report: 1) how the assumptions were verified; 2) standardized regression weights with t-tests and P; 3) unstandardized regression weights with SE, t-tests, *df* and P; 4) R2; and 5) the overall F with *df* and P.

5.2.3. Advanced linear modeling

As mentioned in §5.1, ANOVA analyses can be re-formulated and generalized as multivariate linear regressions. There is a similar hierarchical relationship among more advanced analytic models (Graham, 2008). General linear models (GLM) combine several multivariate linear regression models in a matrix structure. Thus, GLM can incorporate t- and F-tests, ordinary regressions, ANCOVA, and others, and can provide overall and individual tests of the modeled effects. Furthermore, by using link functions to transform non-linear functions into linear functions, GLM can be extended to model non-linear data, thereby accommodating Poisson and other non-linear regressions. Generalized linear mixed models (GLMM) extend GLM to include random effects in the predictor variables.

In structural equation modeling (SEM), factor analysis and multiple regression approaches are combined to analyze relationships between measured and latent variables (Kline, 2015; Schumacker and Lomax, 2016; Ullman and Bentler, 2013). Thus, GLM can be seen as a special case of SEM. SEM can be used to model different connections among predictor variables, latent variables, and measured outcomes. Following specification of a model, SEM estimates its parameters from the data, uses them to generate an estimated population covariance matrix, and estimates the fit of the model from it. Varieties of SEM include exploratory factor analysis and path analysis. SEM also permits estimates of differences in the means of measured and latent variables. Confirmatory factor analysis can be used to test the statistical validity and independence of the constructs included in the SEM, independent of the relationships posited by the model.

Sample sizes for GLM or GLMM depend on the number of variables modeled. Several statistical packages contain algorithms to assess power and sample size (e.g., Williamson, 2020). SEM generally requires quite large samples. Sample size is usually considered adequate if the ratio of sample size n to the number of estimated parameters q, is $n/q \geq 10$, although some argue for $n/q \geq 20$ (Kline, 2015).

5.2.3.1. Assumptions and reporting. These models rest on a number of assumptions. If a maximum likelihood estimation approach is used, the raw or transformed data should be interval or ratio scale data with homoscedastic multivariate normal distributions. Relatively small violations of normality can have large effects on outcomes, so outliers should be censored and missing data imputed. If these assumptions are not met, a least-squares approach should be used. The variables should be linearly related to each other and relatively additive. Error terms should be uncorrelated.

Reporting should include: 1) a diagram of the model with standardized regressions near the arrows and significances indicated; 2) the sample size and ratio of sample size to parameters estimated; 3) the type of matrix analyzed, with link functions identified; 4) the model chi squared, χ^2_M, with *df*, or other overall fit statistic; 5) one or more

approximate fit indices; 6) evidence of linearity or the fit of link functions; 7) unstandardized (with SE) and standardized model parameters; 8) whether there were outliers, and how they were handled. Models with good overall fit should have $\chi^2_M \gg 0.05$; this indicates that the model is consistent with the data, not whether it is actually correct. The most common fit index is the Steiger–Lind root mean square error of approximation (RMSEA), which should be reported with a CI (Steiger, 2010). RMSEA of 0.01, 0.05 and 0.08 are considered excellent, good and poor, respectively. It is considered best practice to compare fits of alternative models and to consider the implications of models with nearly equivalent mathematical fits.

6. Estimation approaches

6.1. Point and interval estimates

The estimation approach is based on estimates of the values of the important experimental outcomes and their precision, i.e., the probability that the estimates fall in a certain range (the confidence interval, CI; typically the 95% CI). These two statistics are usually called point and interval estimates. Often the parameter estimated is the effect size (see §6.2.).

CI for repeated measures designs should be computed separately from those of the individual groups. Blouin and Riopelle (2005) and Masson and Loftus (2003) describe methods.

The estimation approach requires larger sample sizes to function than the statistical significance approach. Cumming (2014) states that if $n < 10$, CI are usually so large as to not be interpretable.

6.2. Effect sizes

Effect sizes are increasingly considered to provide a crucial basis for interpretation (Blume et al., 2019; Betensky, 2019; Cumming, 2012, 2014; Goodman et al. 2019). CI can also be associated with effect sizes (Lee, 2016). For example, for two normally distributed samples of size n_1 and n_2, the effect size estimate is δ (§5.1.9.) and its 95% CI is:

$$\delta - 1.96\ \sigma(\delta) \text{ to } \delta + 1.96\ \sigma(\delta),$$

where $\sigma(\delta) = [\ (n_1+n_2)/n_1n_2 + \delta^2/2\ (n_1+n_2)\ ]^{1/2}$.

Bivariate data have to have bivariate normality to compute an effect size. If the data are not normal, Fisher's z' transform of r can normalize them and permit computation of a 95% CI.

6.3. Power

Estimation approaches do not involve α, so there is no β and statistical power cannot be calculated. Instead, one specifies the size of the maximum CI desired and uses the expected variance of the sample to calculate the sample size required to yield it (Maxwell et al., 2008; Cumming, 2014).

6.4. Multiplicity

Unless the study is exploratory, multiplicity (§5.1.5.) should be controlled if a number of CI are used to analyze a family of hypotheses. Benjamini and Yekutieli (2005) describe an FDR method for this.

6.5. Interpretation and reporting

The interpretation of statistical estimates is introduced in §4.2. CI, effect sizes and many significance testing outcomes are mathematically interconvertible (Altman and Bland (2011) give some examples). The two approaches are, however, epistemologically very different. That is, recognizing that single results are unlikely to be dispositive as to meaning, estimates and their precisions are interpreted in a continuous rather than dichotomous way. The underlying assumptions are [i] that the particular outcome of an experiment is just one of an infinite number of outcomes from the underlying sampling distribution, and [ii] that the best use of the data is in a future meta-analysis. Statistical significance is not assessed, and no particular importance is given to outcomes that would be statistically significant versus outcomes that are similar but would not be significant. Both the upper and lower limits of CI should be discussed. For more discussion, see Cumming (2012, 2014) and Wasserstein et al. (2019).

Point estimates (i.e., the sample means, etc.), interval estimates (e.g., 95% CI or effect sizes), group SD and sample sizes should all be reported. Graphical displays are often effective for reporting CI. The two-dimensional cat's eye representation combines the length of the confidence interval and the shape of its sampling distribution (Cumming, 2014).

7. Nonparametirc statistics

Nonparametric approaches, i.e., those that make few assumptions about the populations sampled (most notably, the assumption of Gaussian population distributions) are used for categorical (nominal)- or ordinal-scale data and for interval- or ratio-scale data that fail to meet the assumptions for parametric tests. For example, because Likert scales (Likert, 1932) are in theory ordinal, non-parametric tests are the safer option for them. Nonparametric tests are generally less susceptible to type-1 errors, but more susceptible to type-2 errors.

7.1. Statistical-significance approach

For categorical (nominal) measurements, the variations on the chi-squared test are usually the best choice: [i] The standard chi-squared tests for differences among expected and observed frequencies in one or more categories related to a single independent variable; [ii] McNemar chi-squared test for differences among expected and observed frequencies if there are paired categories, again with one independent variable; [iii] the Mantel-Haenszel chi-squared for differences among expected and observed frequencies in one or more categories related to two independent variables. These tests break down if the expected or observed frequencies in individual cells are < 6. In this situation, Fisher's exact test can be used.

The chi-squared distribution, upon which the chi-squared test is based, comes up in many contexts. For example, the expected value of sample variances follows the chi-squared distribution. Thus, the F distribution, which is the basis of ANOVA, is the ratio of two chi-squared distributions.

For ordinal (ranked) data, the Mann–Whitney–Wilcoxon test is an appropriate nonparametric version of t-tests for both independent and non-independent samples. It tests for differences in the central tendency (not means) of two groups. This test can be more powerful than the *t*-test if, for example, the data include extreme values. It is important to note that not all computerized statistics packages compute this statistic accurately (Bergmann et al., 2000). The Kruskal-Wallis and Friedman and tests are appropriate nonparametric versions of one-way ANOVA for independent samples and repeated-measures samples of ranked data, respectively.

Nonparametric approaches also require control of multiplicity (see §5.1.5.).

Spearman's rho is an appropriate nonparametric measure of association if one or both variables is an ordinal-scale measurement or the data fail to meet the assumptions of simple regression.

7.1.1. Reporting

Because non-parametric tests use the ordinal structure of the data, central tendency should be reported with medians and ranges, usually the semi-interquartile range or MAD (see §2.8.).

Both the *df* and the sample size are required to specify the probability

level of chi-squared tests, so both should be reported, along with the value of the test statistic and its probability. The Mann–Whitney–Wilcoxon, Kruskal-Wallis and Friedman tests depend only on the group sizes, so these should be reported together with and the test statistics and their probabilities. The significance of Spearman's rho is tested with a *t*-test and reported as described above. As described in §4.2, exact probabilities should be given.

7.2. Estimation approach

Nonparametric estimation methods are not as advanced as the nonparametric significance tests described above, although a number are under development (Brown and Levine, 2007; Powell, 2003, 2003; Soltanian and Hossein, 2012; Wang et al., 2012). Methods based on kernel-density estimation (Parzen, 1962; Rosenblatt, 1956) are used increasingly in both exploratory data analysis (e.g., Harpole et al., 2014) and in analytic statistics (e.g., Miladinovic et al., 2014).

References

Adroher, N. D., Prodinger, B., Fellinghauer, C. S., & Tennant, A. (2018). All metrics are equal, but some metrics are more equal than others: A systematic search and review on the use of the term 'metric. *PLoS One, 13*, Article e0193861.

Alderson, P. (2004). Absence of evidence is not evidence of absence. *BMJ, 328*, 476–477.

Allison, D. B., Paultre, F., Goran, M. I., Poehlman, E. T., & Heymsfield, S. B. (1995). Statistical considerations regarding the use of ratios to adjust data. *International Journal of Obesity and Related Metabolic Disorders, 19*, 644–652.

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *BMJ, 311*, 485.

Altman, D. G., & Bland, J. M. (2011). How to obtain the confidence interval from a P value. *BMJ, 343*, d2090.

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician, 73*(Suppl 1), 262–270.

Anscombe Francis, J. (1973). Graphs in statistical analysis. *The American Statistician, 27*, 17–21.

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biomedical Journal, 52*, 708.

Benjamini, Y., & Hochberg, Y. (1995). The false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57*, 289.

Benjamini, Y., & Yekutieli, D. (2005). Discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association, 100* (71).

Bergmann, R., et al. (2000). Different outcomes of the Wilcoxon—Mann—Whitney test from different statistics packages. *The American Statistician, 54*, 72.

Betensky, R. (2019). The p-Value requires context, not a threshold. *The American Statistician, 73*(Suppl 1), 115–117.

Bier, D. M., Allison, D. B., Alpers, D. H., Astrup, A., Cashman, K. D., Coates, P. M., Fukagawa, N. K., Klurfeld, D. M., Mattes, R. D., & Uauy, R. (2015). Introduction to the series "Best (but oft-forgotten) practices. *American Journal of Clinical Nutrition, 102*, 239–240.

Blouin, D. C., & Riopelle, A. J. (2005). On confidence intervals for within-subjects designs. *Psychological Methods, 10*, 397–412.

Blume, J., Greevy, R., Welty, V., Smith, J., & DuPont, W. (2019). An introduction to second generation p-values. *The American Statistician, 73*(Suppl 1), 157–167.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd). NY, NY USA: Routledge.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.

Boutron, I., Altman, D. G., Moher, D., Schulz, K. F., Ravaud, P., & CONSORT NPT Group. (2017). CONSORT statement for randomized trials of nonpharmacologic treatments: A 2017 update and a CONSORT extension for nonpharmacologic trial abstracts. *Annals of Internal Medicine, 167*, 40–47.

Bramness, J. G., Skurtveit, S., Gustavsen, I., & Mørland, J. (2008). The absence of evidence is not the same as evidence for absence! *Addiction, 103*, 513–514.

Brown, A. W., Kaiser, K. A., & Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Acadamy of Science (USA), 225*, 2563–2570.

Brown, L. D., & Levine, M. (2007). Variance estimation in nonparametric regression via the differences sequence method. *Annals of Statistics, 35*, 2219–2232.

Button, K. S., et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365.

Campbell, N. R. (1920). *Physics: The elements*. Cambridge, UK: Cambridge University Press.

Campbell, M. K., Piaggio, G., Elbourne, D. R., Altman, D. G., & CONSORT Group. (2012). Consort 2010 statement: Extension to cluster randomised trials. *BMJ, 345*, Article e5661.

Carter, R. E. (2013). A standard error: Distinguishing standard deviation from standard error. *Diabetes, 62*, e15.

Caudle, R. M., & Williams, G. M. (1993). The misuse of analysis of variance to detect synergy in combination drug studies. *Pain, 55*(313).

Cohen, J. (1988). *Statistical power Analysis for the behavio ral sciences(2nd ed.)*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychology Review, 57*, 145–158.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Cumming, G. (2014). The new statistics: When and why. *Psychological Science, 25*, 7–29.

Curran-Everett, D. (2000). Multiple comparisons: Philosophies and illustrations. *American Journal of Physiology, 279*, R1.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781.

Dwan, K., Li, T., Altman, D. G., & Elbourne, D. (2019). CONSORT 2010 statement: Extension to randomised crossover trials. *BMJ, 366*, l4378.

Geary, N. (2013). Understanding synergy. *American Journal of Physiology, 304*, E237.

Gelman, A. (2003). Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review, 73*, 369–382.

George, B. J., Beasley, T. M., Brown, A. W., Dawson, J., Dimova, R., Divers, J., Goldsby, T. U., Heo, M., Kaiser, K. A., Keith, S. W., Kim, M. Y., Mehta, T., Oakes, J. M., Skinner, A., Stuart, E., & Allison, D. B. (2016). Common scientific and statistical errors in obesity research. *Obesity, 24*, 781–790.

Gideon, N., Hawkes, N., Mond, J., Saunders, R., Tchanturia, K., & Serpell, L. (2016). Development and psychometric validation of the EDE-QS, a 12 item short form of the Eating Disorder Examination Questionnaire (EDE-Q). *PLoS One, 11*, Article e0152744.

Goodman, W., Spruill, S., & Komaroff, E. (2019). A proposed hybrid effect size plus p-value criterion: Empirical evidence supporting its use. *The American Statistician, 73*, 168–185.

Graham, J. M. (2008). The general linear model as structural equation modeling. *Journal of Educational and Behavioral Statistics, 33*, 485–506.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.

Harpole, J. K., Woods, C. M., Rodebaugh, T. L., Levinson, C. A., & Lenze, E. J. (2014). How bandwidth selection algorithms impact exploratory data analysis using kernel density estimation. *Psychological Methods, 19*, 428–443.

Hartung, J., Cottrell, J. E., & Giffin, J. P. (1983). Absence of evidence is not evidence of absence. *Anesthesiology, 58*, 298–300.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). New York NY USA: Guilford Press.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800.

Ioannidis, J. P. A. (2005). Why most published research finding are false. *PLoS Medicine, 2*, e124.

Ioannidis, J. P. (2014). How to make more published research true. *PLoS Medicine, 11*, Article e1001747.

Ioannidis, J. (2019). What have we (not) learnt from millions of scientific papers with p-values? *The American Statistician, 73*(Suppl 1), 20–25.

Jinbo He, J., Shaojing Sun, S., & Xitao Fan, X. (2021). Validation of the 12-item short form of the eating disorder examination questionnaire in the Chinese context: Confirmatory factor analysis and Rasch analysis. *Eating and Weight Disorders, 26*, 201–209.

Joint. (2008). Joint Committee for Guides in Metrology. Evaluation of measurement data — guide to the expression of uncertainty in measurement. https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.

Kampen, J. (2019). Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables. *Measurement, 137*, 428–434.

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology, 4*, 513.

Kirby, K. N., & Gerlanc, D. (2013). BootES: an R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods, 45*, 905–927.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th Edition). New York, NY USA: Guilford Press.

Kraemer, H. C., & Blasey, C. M. (2004). Centering in regression analyses: A strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research, 13*, 141–151.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement vol 1: Additive and polynomial representations*. San Diego and London: Academic Press.

Kyngdon, A. (2013). Descriptive theories of behaviour may allow for the scientific measurement of psychological attributes. *Theory & Psychology, 23*, 227–250.

Landis, S. C., Amara, S. G., Asadullah, K., Austin, C. P., Blumenstein, R., Bradley, E. W., Crystal, R. G., Darnell, R. B., Ferrante, R. J., Fillit, H., Finkelstein, R., Fisher, M., Gendelman, H. E., Golub, R. M., Goudreau, J. L., Gross, R. A., Gubitz, A. K., Hesterlee, S. E., Howells, D. W., … Silberberg, S. D. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature, 490*, 187–191.

Lee, D. K. (2016). Alternatives to P value: Confidence interval and effect size. *Korean J Anesthesiol, 69*, 555–562.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*, 764–766.

Likert, R. (1932). A Technique for the measurement of attitudes. *Archiv für Psychologie, 140*(1–55).
López Puga, J., Krzywinski, M., & Altman, N. (2015a). Points of significance: Bayes' theorem. *Nature Methods, 12*, 277–278.
López Puga, J., Krzywinski, M., & Altman, N. (2015b). Points of significance: Bayesian statistics. *Nature Methods, 12*, 377–378.
Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*(1–27).
Mair, P., Hatzinger, R., & Maier, M. J. (2017). *Extended Rasch Modeling: The R Package eRm. cran.r-project.org/web/packages/eRm/vignettes/eRm*.
Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 57*, 203–220.
Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample-size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537–563.
Miladinovic, B., Kumar, A., Mhaskar, R., & Djulbegovic, B. (2014). Benchmarks for detecting 'breakthroughs' in clinical trials: Empirical assessment of the probability of large treatment effects using kernel density estimation. *British Medical Journal Open, 4*, Article e005249.
Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534.
Nieuwenhuis, S., et al. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience, 14*, 1105.
Nosek, B. A., Ebersoleb, C. R., DeHavena, A. C., & Mellora, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*, 2600–2606.
Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(1–18).
Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics, 33*, 1065.
Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., … Wurbel, H. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLoS Biology, 18*. e3000410, 2020.
Powell, J. L. (2003a). *Notes on nonparametric density estimation*. Berkely: University of California. http://eml.berkeley.edu/~powell/e241a_sp10/ndnotes.pdf.
Powell, J. L. (2003b). *Master class in semi- and non-parametric econometrics*. Economic and Social Research Council. http://www.cemmap.ac.uk/resource/id/41.
Prnjak, K., Mitchison, D., Griffiths, S., Mond, J., Gideon, N., Serpell, L., & Hay, P. (2020). Further development of the 12-item EDE-QS: Identifying a cut-off for screening purposes. *BMC Psychiatry, 20*, 146.
R Core Team. R. (2021). *A Language and environment for statistical computing*. Wirtschaftsuniversität Wien, Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org/.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests. Danmarks Paedagogiske Institut, Copenhagen, Denmark*. Chicago, IL USA): University of Chicago Press (reprinted 1980.
da Rocha, N. S., Chachamovich, E., de Almeida Fleck, M. P., & Tennant, A. (2013). An introduction to Rasch analysis for psychiatric practice and research. *Journal of Psychiatric Research, 47*, 141–148.
Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics, 27*, 832.
Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88*, 1273.
Sarle, W. S. (1997). *Measurement theory: Frequently asked questions*. Cary NC USA: SAS Institute Version 3. ftp.sas.com/pub/neural/measurement.html.
Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling* (4th ed.). New York, NY USA: Routledge.
Sijtsma, K. (2011). Introduction to the measurement of psychological attributes. *Measurement, 44*, 1209–1219.
Simmons, J. P., et al. (2011). False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359.
Soltanian, A. R., & Hossein, M. (2012). A non-parametric method for hazard rate estimation in acute myocardial infarction patients: Kernel smoothing approach. *Journal of Research in Health Sciences, 12*, 19–24.
Steiger, J. (2010). Structural model evaluation and modification - an interval estimation approach. *Multivariate Behavioral Research, 25*, 173–180.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.
Stevens, J. (2001). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ USA: Lawrence Erlbaum Associates.
Tal E. "Measurement in Science", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2020/entries/measurement-science/>.
Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
Ullman, J. B., & Bentler, P. M. (2013). Structural equation modeling. In I. B. Weiner (Ed.), *Handbook of psychology* (2nd ed., pp. 661–690). Hoboken, NJ USA: John Wiley & Sons.
Wainer, H. (2007). *Visual revelations*. New York, NY: Copernicus – Springer.
Wainer, H. (2008). Velleman. P. Looking at blood sugar. *Chance, 21*, 56–61.
Wang, Q., Dinse, G. E., & Liu, C. (2012). Hazard function estimation with cause-of-death data missing at random. *Annals of the Institute of Statistical Mathematics, 64*, 415–438.
Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician, 73*(Suppl 1), 1–19.
Wilcox, R. (2003). *Applying contemporary statistical techniques*. Houston, TX: Elsevier: Gulf Professional Publishing.
Williamson, M. (2020). Sample size calculation with R: Generalized linear mixed models. Dakota cancer collaborative on translational activity. https://med.und.edu/daccota/_files/pdfs/berdc_resource_pdfs/sample_size_r_module_glmm2.pdf.
Winer, B. J. (1971). *Statistical principles in experimental design*. New York NY: McGraw-Hill.

Nori Geary [*,1]
Department of Psychiatry, Weill Medical College of Cornell University, New York, NY, USA

Suzanne Higgs
School of Psychology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[*] Corresponding author.
E-mail address: nori.geary@gmail.com (N. Geary).

[1] Retired.