General Article

Advances in Methods and Practices in Psychological Science
July-September 2021, Vol. 4, No. 3, pp. 1–18

DOI: 10.1177/25152459211035109
www.psychologicalscience.org/AMPPS

Summary Plots With Adjusted Error Bars: The *superb* Framework With an Implementation in R

Denis Cousineau[1], Marc-André Goulet[1], and Bradley Harding[2]
[1]École de psychologie, Université d'Ottawa, Ottawa, Canada, and [2]École de psychologie, Université de Moncton, Moncton, Canada

Abstract

Plotting the data of an experiment allows researchers to illustrate the main results of a study, show effect sizes, compare conditions, and guide interpretations. To achieve all this, it is necessary to show point estimates of the results and their precision using error bars. Often, and potentially unbeknownst to them, researchers use a type of error bars—the confidence intervals—that convey limited information. For instance, confidence intervals do not allow comparing results (a) between groups, (b) between repeated measures, (c) when participants are sampled in clusters, and (d) when the population size is finite. The use of such *stand-alone* error bars can lead to discrepancies between the plot's display and the conclusions derived from statistical tests. To overcome this problem, we propose to generalize the precision of the results (the confidence intervals) by adjusting them so that they take into account the experimental design and the sampling methodology. Unfortunately, most software dedicated to statistical analyses do not offer options to adjust error bars. As a solution, we developed an open-access, open-source library for R—*superb*—that allows users to create summary plots with easily adjusted error bars.

Keywords

error bars, confidence interval, plots, results, statistics, standard error, within-subjects designs, sampling method, research methods, open materials

Received 12/19/19; Revision accepted 7/2/21

After gathering the necessary data from an experiment, researchers are likely to plot their results before interpreting them. Although plots can be used to display raw data (e.g., scatterplots), in this article, we focus on plots that display summary statistics (e.g., mean plots). To be informative, summary plots must convey some indications of the results' uncertainty. To that end, all summary statistics should be accompanied by error bars indicating the precision with which such results were obtained (Cumming, 2014; Loftus, 1993; Wilkinson & The Task Force on Statistical Inference, 1999).

Sometimes, error bars represent the standard deviation of the sample. However, many authors have recommended using bars depicting a range of values relevant to assess the true population parameter, either the standard error or the confidence interval (Cumming & Fidler, 2009; Cumming & Finch, 2001, 2005). In the long run, the former has about 68% chance of containing the population mean (if the population is normally distributed), whereas for the latter, the coverage is up to the researcher, and 95% is common. All descriptive statistics have an associated measure of standard error or confidence interval (for reviews, see Goulet-Pelletier & Cousineau, 2018; Harding et al., 2014, 2015). For concision's sake, we concentrate herein on the mean and its confidence interval, but everything that follows is equally applicable to other descriptive statistics, such as the median or skewness. We return to this in the Discussion.

Corresponding Author:
Denis Cousineau, École de psychologie, Université d'Ottawa
Email: denis.cousineau@uottawa.ca

The standard error of the mean (SE_M) is estimated with

$$SE_M = S/\sqrt{n}, \quad (1)$$

in which S is the sample standard deviation and n is the sample size. Transforming a standard error of the mean into a confidence interval of the mean (CI_M) requires increasing the coverage of this interval to a desired level. For the mean, the t distribution is used to get a coverage factor, t_c, given the desired coverage probability and the degree of freedom, such that

$$CI_M = M \pm SE_M \times t_c. \quad (2)$$

Equations 1 and 2 are well known. What is less known is that the resulting error bars offer limited information when comparing two groups and when comparing repeated measures, among others. These error bars should be understood as *stand-alone* bars: They apply to a single mean examined in isolation (the equivalent of a one-sample t test). Unfortunately, it is very uncommon to perform an experiment with a single group measured in a single condition. Hence, most of the time, the stand-alone error bars do not match the intentions of the researchers using them.

Through the years, a number of adjustments to the stand-alone measures of precision have been proposed, starting with the seminal work of Loftus and Masson (1994) for comparing means in within-subjects designs and Goldstein and Healy (1995) for comparing means in between-subjects designs (also see Baguley, 2012; Cousineau, 2019; Franz & Loftus, 2012; Nathoo et al., 2018; Tryon, 2001). Later, other adjustments were found to compare means from data that were not sampled randomly (Cousineau & Laurencelle, 2016) or coming from finite-size populations (Thompson, 2012).

When performing statistical tests of means, the procedure is adapted to the experimental design. For example, a researcher would use a two-group t test when comparing two groups or a paired-sample t test when comparing two repeated measures. These tests are tailored to the experimental design and to the purpose of comparing means. In that case, why would the stand-alone confidence intervals be used in all these diverse situations? Using these unadjusted confidence intervals is the equivalent of using a single-group t test. Would it be adequate to use that test in all situations? The answer is certainly not. Yet in making plots, this is exactly what is being done.

The purpose of this article is to propose and provide a coherent and unified approach to the computation of error bars for a wide array of situations. We call this the *superb* (summary plots with adjusted error bars) framework. In what follows, we first review the limitations of stand-alone confidence intervals. We then present the four classes of adjustments that compose *superb*. To demonstrate the viability of this framework, it was programmed into an R library that provides all the necessary adjustments by selecting options, as described in Appendix A. Before concluding, we examine the relation of these adjustments with well-known statistical tests of means. However, in the Discussion, we open the framework to a much wider array of utilization.

What Really Is a Stand-Alone Confidence Interval?

The stand-alone confidence interval of a parameter (e.g., the population mean) is an estimator of an interval that—over many samples—has a certain probability of containing the true population parameter. The probability, chosen by the researcher, is called the *confidence level* (often noted with γ). Technically, it estimates the quantiles of the predictive distribution of the population parameter (assumed to be a constant) given an observed sample (Rouanet & Lecoutre, 1983). It is an equal-tail interval in the sense that both extremities have an equal, weak chance of containing the population parameter in the long run. Confidence intervals are related to statistical tests: When the hypothesized population parameter is not included in the interval, the test will reject that hypothesis with a p value below $\alpha = 1 - \gamma$ (Greenland et al., 2016). For instance, a 95% confidence interval contains all the values for which $p > .05$. We call it a *stand-alone confidence interval* because a single sample is all that is required to assess the significance of a single parameter.

The stand-alone confidence interval of the mean of Equation 2 is based on assumptions. Most notably, it assumes the normality of the population. This assumption may be problematic (and we consider a workaround in the Discussion). However, it also surmises four additional assumptions. First, it assumes that the interval is used to compare a result with a fixed value (e.g., a value posed in a hypothesis). Yet almost always, researchers want to compare conditions with other conditions whose values are not preestablished. When comparing two samples, their mean separation is a different parameter, and its precision, being based on two imprecise means, is weaker (Cousineau, 2020). Thus, in this situation, the stand-alone confidence intervals of the mean are too short. Second, even when confidence intervals are adapted to relative assessments, they also assume that groups are measured independently. Yet repeated measures (e.g., pre-post designs) are frequently used because they are generally statistically more powerful than independent-sample designs. Increasing statistical

power translates into improved precision. Consequently, the stand-alone confidence intervals of the mean are likely too long in such designs. Third, the data sampling method is assumed to be *simple randomized sampling*. This is not the only possible method, and other sampling techniques are often found in the literature that may increase or decrease precision (e.g., stratified sampling in survey studies or cluster randomized sampling in education; more on this later). When cluster randomized sampling is used, the stand-alone confidence intervals of the mean are probably too short. Finally, it assumes that the population size is very large or infinite. Again, for studies examining specific populations, the population size might be finite. Sampling a sizable part of a finite population improves precision. Thus, when the population size is small, the stand-alone confidence intervals of the mean are too long.

Past authors have published heuristics to handle some of these methodological considerations (beginning with Cumming & Finch, 2005). However, it was on the reader's onus to mentally make these adjustments with all the approximations and biases that they entail. Yet the impact of such methodological considerations on precision is known with exactitude. Therefore, error bars can be adjusted to exactly reflect the experimental situation. The purpose of the *superb* framework is to provide these adjustments so that confidence intervals adequately reflect the research's design and objectives.

As will be illustrated with the examples that follow, the risk with stand-alone confidence intervals is that researchers provide, possibly unbeknownst to them, visual information that contradicts the inferences made with statistical tests (Estes, 1997; Goldstein & Healy, 1995; Loftus & Masson, 1994). It would be unfortunate that such contradictory information plants the seeds of suspicion in a reader's mind. The risk is larger when the reader is unaware of the existence of heuristics to interpret stand-alone confidence intervals and could have the detrimental effect that the readers no longer consider error bars. The present framework does not require heuristics from the readers, and all confidence intervals are to be interpreted in the exact same manner, which encourages their use.

The *superb* Framework

The *superb* framework implements the principle that adequate error bars must support (and consequently can complement) results obtained from other techniques. In the present case, the confidence intervals are like inverted statistical inference because they contain all results for which $p > \alpha$. Following Tryon (2001), we may call the adjusted confidence intervals *inferential intervals*. To that end, the stand-alone confidence intervals undergo a series of adjustments to incorporate additional information with regard to the experimental design and the sampling methodology, all of which are known to alter precision. Each adjustment is applied to correct the length of the error bars by using a multiplicative correction factor; all adjustments can be combined by multiplying the correction factors. Thus, adjusting error bars is simple to implement and easy to understand. For example, a correction factor smaller than 1 implies shorter confidence intervals and, consequently, a data set with improved precision in comparison with the stand-alone confidence intervals that would have initially, and erroneously, been used.

Four classes of adjustments exist to account for these various methodological situations. The first adjustment accounts for the purpose of the experiment, the second accounts for the experimental design; the last two are adjustments for the sampling methodology and the population size. Each adjustment is described and illustrated with a brief example. The data and the script used to construct the figures are available at https://dcousin3.github.io/superb/articles/TheMakingOf.html. All the data are fictitious.

Adjustment for the purpose of the research

When the purpose of the experiment is to compare a single result with a prespecified value (i.e., is the average IQ of a certain group different from 100?), the stand-alone confidence interval of the mean is correct because it is designed precisely for that purpose. However, in most studies, means are to be compared with other means to assess their difference. When this is the case, an adjustment must be made to account for the additional uncertainty regarding the position of the two means relative to one another.

The simplest difference adjustment is a multiplier of $\sqrt{2}$, which must be applied to the length of the confidence intervals to account for the additional variability in the relative location of the two means that are compared (Baguley, 2012; Estes, 1997; Franz & Loftus, 2012; Goldstein & Healy, 1995; Pfister & Janczyk, 2013). This adjustment assumes that the groups have homogeneous variances. Because $\sqrt{2} = 1.41$, the error bar lengths are increased by 41% under this adjustment. As discussed in Baguley (2012), this adjusted confidence interval is not about the exact position of a mean; it is about the lag between two means.

When the variances are heterogeneous, an alternative is the Tryon adjustment (Tryon 2001). Given the variances and the sample sizes in each group, it returns an adjustment that is 1.41 or larger (although it is rarely larger than 1.75). We return to this adjustment later.

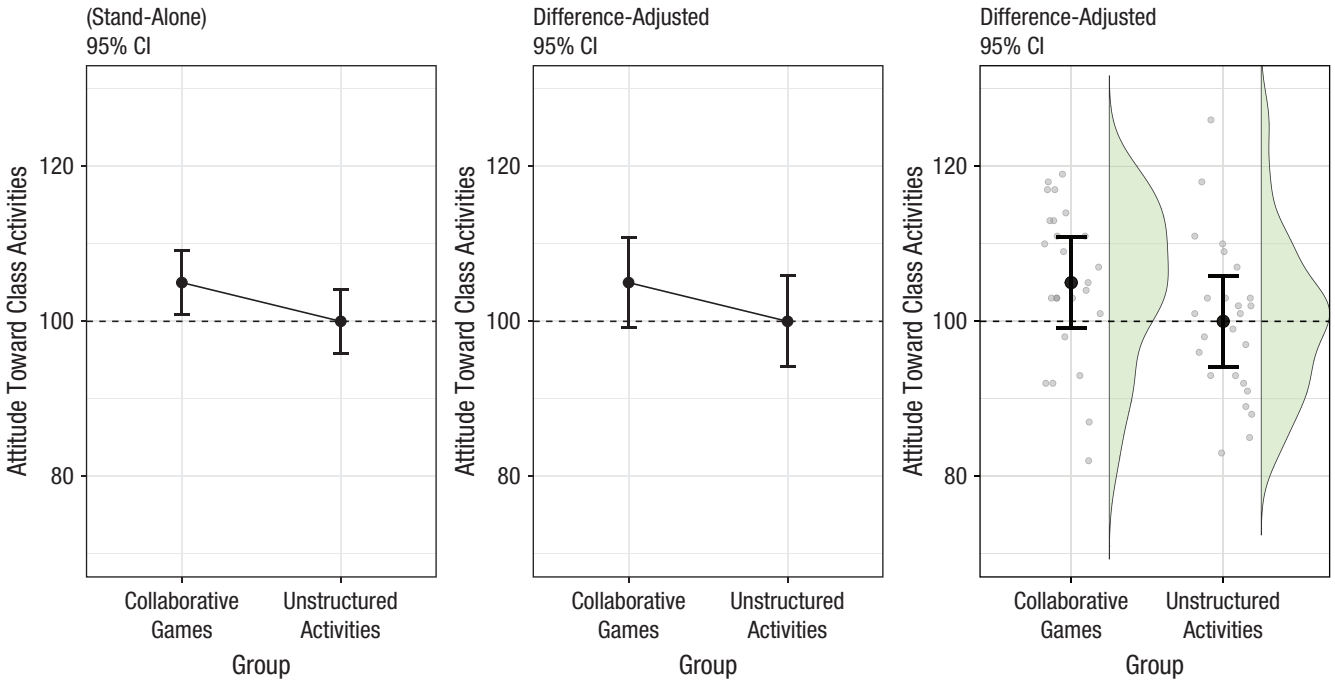

Fig. 1. Mean scores of fictional data examining attitude toward school activity. The dashed line corresponds to the mean of the control group. The left plot shows stand-alone 95% confidence intervals of the mean. Here, the mean of the control group is not included within the error bar of the experimental group mean, which suggests a significant difference between the two groups. In the central panel are the difference-adjusted 95% confidence intervals of the mean. Here, the mean of the control group is included within the error bar of the experimental group, which indicates rightly a nonsignificant difference between the two groups. The right panel shows a raincloud plot in addition to the difference-adjusted 95% confidence intervals of the mean.

Example 1. Consider a study examining the attitude of primary school students toward engaging in class activities. The participants are divided into two groups, one performing collaborative games when arriving to school and the other engaging in unstructured activities.

When comparing the two groups' data with a *t* test, we find that the difference is not statistically significant, $t(48) = 1.76$, $p = .085$. Yet the stand-alone 95% confidence intervals of the mean seen in Figure 1 (left), suggest otherwise: The mean of the second group is not included in the first group interval (and vice versa). Applying a correction for the purpose of comparing the two groups, we get the plot in the center of Figure 1, which shows that the difference-adjusted 95% confidence intervals of the mean include the mean of the other condition, which suggests—rightly so—a lack of statistical difference at the .05 (1 − .95) level. The 41% increase in confidence interval lengths is enough to correct a situation that if gone unadjusted, would have returned conflicting information regarding the statistical relationship between both groups (statistical tests vs. plot).

If instead of using the difference-adjusted confidence intervals, the plot had been based on the Tryon-adjusted confidence intervals, the change would have been unnoticeable; the two correction factors are actually identical to three decimals because the two groups' standard deviations are nearly identical (10.1 vs. 9.9).

In the third panel of Figure 1, a raincloud layout is used that shows the raw data with jittered dots and their distribution with a half-violin plot (Allen et al., 2019; Lane, 2019; Marmolejo-Ramos & Matsunaga, 2009; Rousselet et al., 2017; Weissgerber et al., 2015; Yang et al., 2021).

Adjustment for the experimental design

When comparisons are made between repeated measures, the confidence intervals may be adjusted further to account for the correlation between the data. For example, the correlation may arise because high-scoring participants tend to produce high scores on all measurements and vice versa. A simple adjustment, the correlation adjustment (CA), consists of multiplying the length of the error bars by the square root of 1 minus the average correlation between the measurements (Cousineau, 2019). Such confidence intervals are to be appraised relative to the lag between means and apply only to means of scores that possess a correlation similar to the observed correlation. They therefore are meaningless in different research contexts that have different correlations and in different experimental designs (e.g., with independent groups). In other words, they are tailored specifically for the present experiment. This is analogous to the fact that we used statistical tests adapted to the present experiment.

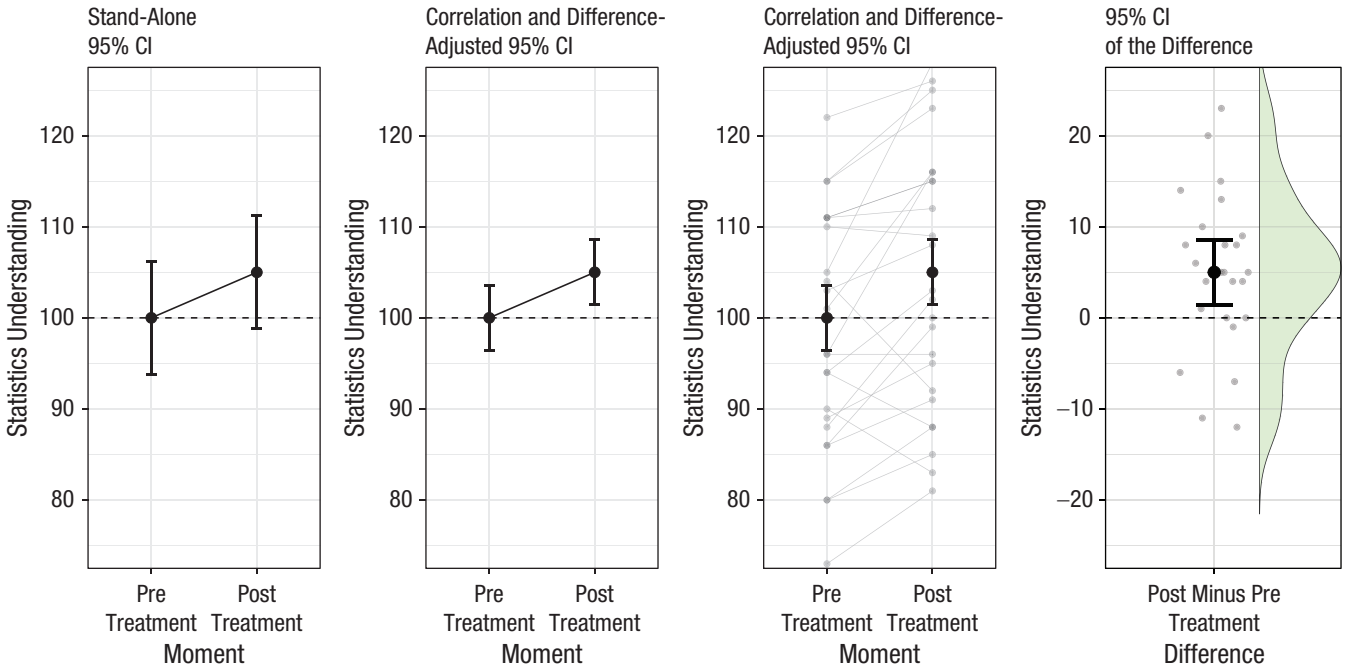

Fig. 2. Effect of the visualization exercises on statistics understanding. The dashed line corresponds to the mean of the pretest measurement. The leftmost plot shows stand-alone 95% confidence intervals of the mean. Here, the mean of the posttreatment measurement is included within the error bars of the pretreatment mean, which suggests a nonsignificant difference. The second plot shows correlation- and difference-adjusted 95% confidence intervals of the mean. Here, the mean of the posttreatment measurement is not included within the error bars of the pretreatment mean, which suggests rightly a significant difference between the two measurements. In the third panel, each participant's data are connected with a line, which shows a benefit of the treatment for a strong majority of the participants. In the last panel, only the difference scores are shown.

This CA is based on the assumption that the data are compound symmetric, a stronger assumption than sphericity.[1] Because the correlation is a pairwise statistic, it necessarily implies that the measures are compared. Therefore, the difference adjustment must also be applied in conjunction with the correlation adjustment.

When compound symmetry is not met, there is no multiplicative adjustment. However, some authors have suggested decorrelating the raw data using transformations before computing precision (Baguley, 2012; Bakeman & McArthur, 1996; Cousineau, 2005, 2017; Cousineau & O'Brien, 2014; Loftus & Masson, 1994; Morey, 2008). See https://dcousin3.github.io/superb/articles/Vignette8.html for definitions of these transformations. There exist two main techniques to obtain decorrelated data, one that gives all data points the same pooled standard deviation (referred to hereafter as *LM*, which stands for Loftus & Masson, 1994) and one that preserves the different standard deviations for each measurement (referred to hereafter as *CM*, which stands for Cousineau-Morey, following Baguley, 2012).

Example 2. Consider the pretreatment and posttreatment scores of 25 college students invited to perform visualization exercises to see whether it would increase statistics understanding. The design is a within-subjects design with two measurements.

Performing a paired-sample *t* test, we get a statistically significant difference, $t(24) = -2.91$, $p = .008$. The difference is 5.00 points, with a standard error of the difference of 1.72; the 95% confidence interval of the difference is [1.45, 8.55].

Despite this significant result, note that the standalone confidence intervals (Fig. 2, left) suggest a lack of significant difference. The adjusted confidence intervals seen on the center of Figure 2 rightly indicate the statistically significant difference.

In the central panel of Figure 2, a correction for correlation was used (the CA method; but the CM or LM methods would do equally well). It reduced the length of the confidence intervals. In the present data, the correlation is .84 (which may be high for human data; see Goulet & Cousineau, 2019). Consequently, the confidence interval lengths, which were first increased by a factor of 1.41 for the difference adjustment, are subsequently shortened by a factor of $\sqrt{1-.84} = 0.40$. Relative to the standalone confidence intervals of the first panel in Figure 2, it represents a net reduction of 44%.

The third panel of Figure 2 further shows each participant using a line. This makes it possible to see

whether there is a trend (upward here) at the level of each participant. This trend is absent or reversed only for five participants out of 25. The fourth and last panel shows only the difference scores. In this last panel, there is no adjustment because the mean difference is to be compared with zero, a fixed value.

Adjustment for the sampling design

The standard for sampling participants from a population is called s*imple randomized sampling* (SRS). It requires each member of the population to have an equal chance of being part of the study. SRS is the reference and requires no adjustments to the confidence interval of the mean.

However, it is not always possible to implement SRS. An alternative sampling method is *cluster randomized sampling* (CRS). In such methodology, clusters of participants are selected randomly, but all the members of the cluster take part in the study. CRS is often used in education, in which the clusters are classes.

In a situation in which the participants are recruited in clusters, there is a possibility that the cluster members are more homogeneous (compared with participants recruited randomly) because of a common history/environment. When this is the case, a small number of clusters reduces the chance of seeing the entire population's variance, and consequently, statistical procedures may be too liberal for an astute interpretation of significance. An adjustment for CRS uses the number of clusters and the number of participants per cluster in addition to an intraclass measure of correlation (Shrout & Fleiss, 1979). The resulting error bars are most commonly increased under CRS. The adjustment is given by a correction factor called λ (Cousineau & Laurencelle, 2016). Again, the cluster-adjusted confidence intervals are meaningful in this experimental context only and should not be used to make inference relative to means obtained in different experimental situations.

Another sampling alternative is the *stratified sampling* method, by which the participants are selected to preserve some characteristics of the population, for example the prevalence of each age category. There is no known solution to adjust the confidence interval of the mean (it is even not known whether confidence intervals are too short or too long and which parameters are relevant to assess this; Kish, 1965; Thompson, 2012).

Example 3. The study examined the policymaking of teams composed of risk managers from various governmental agencies under a catastrophe scenario. The teams could either include members that have similar higher education and the same field of expertise (*homogeneous teams*) or include members that have various levels of education and different fields of expertise (*heterogeneous teams*). Although the teams were randomly selected, all its members participated in the study, which therefore resulted in a cluster-randomized sample. In this example, all the teams contain five members, but there can be any number of members per cluster.

Figure 3 (left) shows the stand-alone 95% confidence intervals (unadjusted for the presence of clusters of participants in the data). They suggest a significant difference between the two groups. However, a hierarchical linear model analysis indicates the lack of a statistically significant difference, $t(4) = 1.42$, $p = .11$ (Bryk & Raudenbush, 1992). When adjusted for the presence of clusters (Fig. 3, center), we get the correct picture. The λ parameters in each group equal 1.80 for Group 1 and 1.36 for Group 2. Along with the difference adjustments (1.41), the confidence intervals are expanded by factors of 2.5 and 1.9, respectively. Fig. 3 (right) shows a raincloud layout (however, because the cluster members are not visible, it is not clear how useful that additional information is).

Adjustment for the population size

A final adjustment can be made if the size of the population is finite and known and if the sample is large enough that it represents a sizable proportion of said population (generally more than 5% of the population). When the population is limited in size, any additional participant makes the sample closer to the whole population, which allows for more precise inferences. If the sample contains more than 5% of the population, researchers are encouraged to use the finite-population adjustment (Cochran, 1953; Kish, 1965; Thompson, 2012). This adjustment is given by the square root of the proportion of the population left unsampled.

Example 4. Consider a sample in which 25 patients suffering from Hutchinson-Gilford-Progeria syndrome are examined. There is, at this time, only 50 patients known with this illness still alive. Hence, the current sample is actually very large, containing 50% of the whole population.

In Figure 4 (left), the stand-alone confidence interval suggests no improvement following a treatment relative to the baseline metabolic value of 100. Yet informing the interval that the whole population is composed of 50 individuals, the correct t test (Kish, 1965) returns a statistically significant difference to 100, $t(24) = 2.64$, $p = .007$. The population-size-adjusted confidence interval plotted in the Figure 4 (center) is congruent with this result. The adjustment is $\sqrt{.50} = 0.71$ so that the error bar is reduced by almost one third (29%). Figure 4 (right) shows a layout in which a violin plot is visible in addition to the summary results. In this example, the

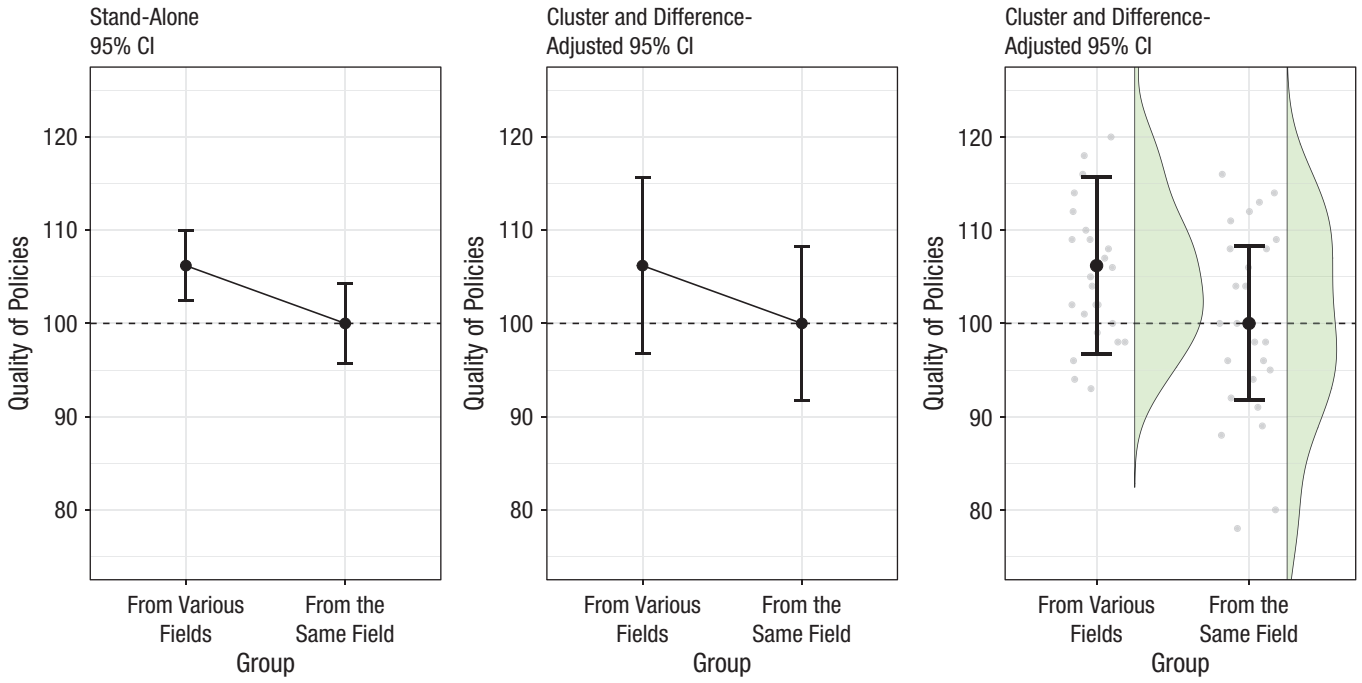

Fig. 3. Mean scores of homogeneous and heterogeneous teams on the quality of risk management policies. The dashed line corresponds to the mean of the second group. The leftmost plot shows stand-alone 95% confidence intervals of the mean. Here, the mean of the control group is not included within the error bars of the experimental group mean, which suggests a significant difference between the two groups. The center plot shows cluster- and difference-adjusted 95% confidence intervals of the mean. Here, the mean of the control group is included within the error bars of the experimental mean, which suggests rightly a nonsignificant difference between the two groups. The third panel adds a raincloud plot to the second.

difference adjustment is not used because there is a single group that is compared with a fixed value.

Naming the adjustments

To be interpretable, it is necessary that error bars be identified clearly and indicate which adjustments were made. There is no official nomenclature as of now, but some terms have been proposed to unify the practice (Baguley, 2012). Here, we suggest labels to unambiguously identify all adjustments described so far. These labels are listed in Table 1. With these labels, the expression *difference-adjusted 95% confidence intervals* would denote intervals proper for comparisons, whereas *pooled correlation- and difference-adjusted 95% confidence intervals* represents Loftus and Masson's (1994) proposal in which all the error bars of the within-subjects factors are of the same length (*pooled*) and based on decorrelated data (*correlation-adjusted*). To our knowledge, no author has ever proposed to pool error bars across groups, and so this feature is absent from *superb*. Finally, the stand-alone confidence intervals—those obtained from most computer software—are plainly termed *95% confidence intervals*. The adjustments need be indicated only in the figure caption to avoid overloading the main text.

To conclude this section, in Appendix A, we provide a quick overview of an R library that implements the *superb* framework, the `superb` library, and in Appendix B, we review the various computations underlying these adjustments.

Relation to Statistical Tests of Mean and the Golden Rule of Adjusted Confidence Intervals

In the *superb* framework, the confidence intervals of the means are adjusted to account for the purpose of the study (comparison or examination in isolation), for the experimental design (independent groups or repeated measures), for the sampling methodology used, and for situations in which the population is of finite size.

Herein, we briefly show that all the adjustments are exact, updating the standard errors from which the confidence intervals are obtained. To that end, we use the *t* test as a reference point. Keep in mind that the difference-adjusted confidence intervals are to be appraised with respect to the lag between means in a specific experimental design. It therefore cannot be interpreted in isolation or with means obtained using different experimental designs.

In a graphical representation, inference is made by examining whether one mean is included in the other mean's confidence interval or, equivalently, whether the lag between the two means, $|M_1 - M_2|$, is smaller than

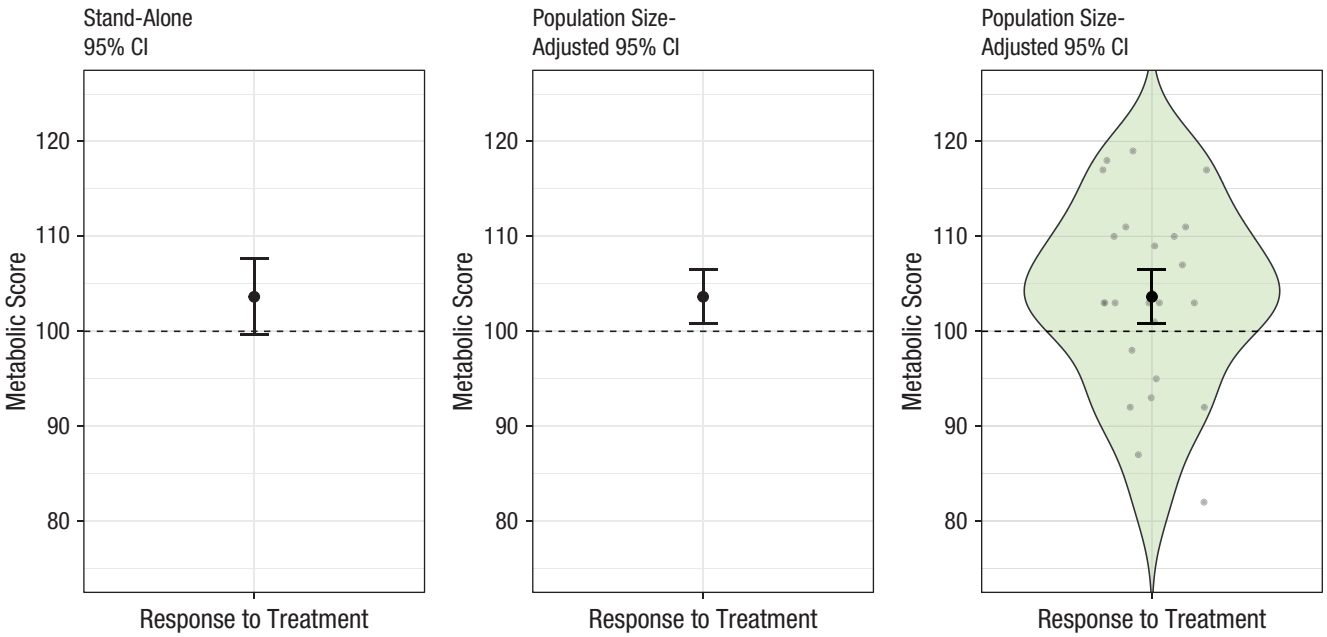

Fig. 4. Mean scores of patients treated. The dashed line corresponds to the baseline score. The leftmost plot shows a stand-alone 95% confidence interval of the mean. Here, the baseline is included within the error bar, which suggests a nonsignificant difference. The center plot shows a population-size-adjusted 95% confidence interval of the mean, which suggests a significant improvement. In the right panel, a violin plot is added to the adjusted confidence interval picturing the distribution of the sample, shown with jittered dots.

the confidence interval's length. When the groups have homogeneous standard deviations and equal sample sizes (as was the case in the previous examples), all the bars' lengths are nearly identical. By examining inclusion, we check that $|M_1 - M_2|$ is smaller than the difference-adjusted confidence interval length. In formula, we get

$$|M_1 - M_2| < \sqrt{2}\, t_c\, S/\sqrt{n},$$

which is equivalent in mathematical form to

$$\frac{|M_1 - M_2|}{\sqrt{2}\, S/\sqrt{n}} < t_c.$$

This is exactly a t test in which S is estimated with the pooled standard deviation. The only difference is that the coverage t_c in the plot is based on separate degrees of freedom ($n-1$), whereas the t test pools the degrees of freedom together, $2(n-1)$.

When the standard deviations, the sample sizes, or both are heterogeneous, the bars will likely be of differing lengths. In that case, the average confidence interval length is the reference. There are two ways to average confidence intervals.

Average in the square sense. Technically, standard deviations cannot be added and, consequently, cannot be averaged. Their square, the error variance, can and so is the squared lengths of confidence intervals. Thus, confidence

Table 1. Suggested Names Given to the Confidence Intervals Adjustments

Type of adjustments	Label	Equation
Experimental design adjustments		
Purpose	*Difference-adjusted* *Tryon-adjusted*	B.5
Repeated measures	*Correlation-adjusted*	B.1 & B.2 or B.6
Sampling design adjustments		
Cluster randomized sampling	*Cluster-adjusted*	B.7
Finite population size	*Population-size-adjusted*	B.8
Make the lengths all equal by		
Pooling the standard deviations	*Pooled*	B.3

Note: The equation numbers refer to Appendix B.

intervals must first be squared, averaged, and then the result squared root (keep with us, Solution 2 is much easier). Computing the mean in the square sense, noted Mean* for short, we get

$$\begin{aligned}
&\text{Mean}^*(\text{ two difference-adjusted CI})\\
&= \text{Mean}^*\left(\sqrt{2}\ t_c\ SE_1, \sqrt{2}\ t_c\ SE_2\right)\\
&= \sqrt{\text{Mean}\left((\sqrt{2}\ t_c\ SE_1)^2, (\sqrt{2}\ t_c\ SE_2)^2\right)}\\
&= \sqrt{\text{Mean}\left(2\ t_c^2\ SE_1^2, 2\ t_c^2\ SE_2^2\right)}\\
&= \sqrt{2}\ t_c\ \sqrt{\text{Mean}(SE_1^2, SE_2^2)}\\
&= \sqrt{2}\ t_c\ \sqrt{\frac{SE_1^2 + SE_2^2}{2}}\\
&= t_c\ \sqrt{SE_1^2 + SE_2^2},
\end{aligned}$$

in which SE_i denotes the standard error of the *i*th mean (Equation 1). Hence, when visually comparing lags to the average length, we actually do reject equality

$$\begin{aligned}
&\text{if } |M_1 - M_2| > \text{Mean}^*(\text{two difference-adjusted CI})\\
&\text{if } |M_1 - M_2| > t_c\ \sqrt{SE_1^2 + SE_2^2}\\
&\text{if } \frac{|M_1 - M_2|}{\sqrt{SE_1^2 + SE_2^2}} > t_c,
\end{aligned}$$

which is exactly a Welch test (Derrick et al., 2016; Welch, 1938). In other words, a Welch test is obtained any time separate confidence intervals are averaged in the square sense. This test is more general than the two-group *t* test, but including this test as a special case should be preferred at all times (Delacre et al., 2018).

Here again, the graphical confidence intervals have separate degrees of freedom ($n_i - 1$, in which n_i is the number of participants in the *i*th group), whereas the Welch test suggests a rectified degree of freedom that can be $n_1 + n_2 - 2$ or smaller when heterogeneity is large. See https://dcousin3.github.io/superb/articles/Vignette7.html to see how to compute Welch degrees of freedom and how to use custom degrees of freedom in the confidence intervals plotted by *superb*. Note that as soon as the samples are not trivially small, the difference between (default) unpooled degrees of freedom and rectified degree of freedom are small so that visually, it is impossible to determine which were used.

Simple average. The mean in the square sense returns a length similar to the mean in the usual sense. It is a little bit more influenced by long lengths than by short lengths so that it tends to be slightly longer than the regular mean. If this bias is to be avoided, an alternative is to use the Tryon adjustment (Tryon, 2001). This adjustment replaces the standard errors (which cannot be averaged in the usual sense) by slightly longer standard errors. By doing so, the usual mean on these altered lengths results in the correct length required for comparisons. Tryon's adjustment, $2 \times E$, is based on the quantity E defined in Tryon (2001; see Appendix B). In the special case in which both sample sizes and variances are homogeneous, it equals $\sqrt{2}$ as the difference adjustment. When this is not the case, Tryon's adjustment returns an adjustment a little larger than 1.41 to compensate for the underestimation introduced by regular averaging. With this adjustment, the means in the regular sense of the Tryon-adjusted confidence interval are

$$\begin{aligned}
&\text{Mean(two Tryon-adjusted CI)}\\
&= \text{Mean}\left(2\ E\ t_c\ SE_1, 2\ E\ t_c\ SE_2\right)\\
&= 2\ E\ t_c\ \text{Mean}(SE_1, SE_2)\\
&= 2\ E\ t_c\ \frac{SE_1 + SE_2}{2}\\
&= 2\ \frac{\sqrt{SE_1^2 + SE_2^2}}{\sqrt{SE_1^2} + \sqrt{SE_2^2}}\ t_c\ \frac{SE_1 + SE_2}{2}\\
&= 2\ \frac{\sqrt{SE_1^2 + SE_2^2}}{SE_1 + SE_2}\ t_c\ \frac{SE_1 + SE_2}{2}\\
&= t_c\ \sqrt{SE_1^2 + SE_2^2}.
\end{aligned}$$

Hence, with Tryon-adjusted confidence intervals, a Welch test is performed visually with a regular mean of the error bar lengths.

As we show, all the ingredients of a Welch test are visually accessible in a plot with separate precision estimates. Furthermore, Tryon adjustment makes it easier to average the two error bar lengths. Finally, it shows that with unequal error bars, inference must be based on an average of the bars.

Similar formal derivations can be done for the CA (Cousineau, 2019; Cousineau & Goulet-Pelletier, 2021), for the cluster adjustment (Cousineau & Laurencelle, 2016), and for the population-size adjustment (Thompson, 2012). All are based on whether the lag between means is shorter than the adjusted confidence interval length (or their average).

Thus, any plot showing the adequately adjusted confidence intervals can be interpreted using a unique rule, the *golden rule of adjusted confidence intervals*:

> If a mean is not included within the confidence interval of another mean, it can be interpreted as being statistically different from that mean at the

$1 - \gamma$ level. When adjusted confidence intervals are of unequal lengths, difference-adjusted ones must be averaged in the square sense, whereas Tryon-adjusted ones must be averaged in the usual sense.

The key in this rule is the notion of *inclusion*. There have been heuristics published in the past that focused on *overlap*; however, this metric is inappropriate for adjusted confidence intervals.

The golden rule is mathematically equivalent to an unpooled Welsh test of mean. However, its application in plots is approximate in two ways. First, the exact limits of the confidence intervals are difficult to position visually precisely because the markers indicating the end of the confidence intervals have a certain thickness and it is not clear whether that thickness is located outside the interval, entirely within the interval, or partially inside and outside of the interval. Second, by default, the confidence intervals are based on separate degrees of freedom, whereas the statistical tests pool all the degrees of freedom together (in the R package, custom degrees of freedom can be provided to override the ones generated by default).

We can estimate the impact of these sources of approximation. Let us assume that we examine two groups with 20 observations per groups. The 95% confidence interval would be based on a critical t value of 2.02. If the limits are visually assessed with ±2% error and based on unpooled degrees of freedom, the resulting confidence interval could be inferred to be based on a critical t value of 2.13 and when perceived too short, to be based on a critical t value of 1.93. This is, in both cases, a 5% misestimation of the bar's length; they correspond to a 96.0% and a 94.6% confidence interval. Hence, for small samples, misestimation results in a confidence level within ±1% of the desired 95% confidence level; for large samples, the confidence level is within ±0.5% of the desired level. Hence, if Rosnow and Rosenthal (1989, p. 1277) are right to say, "Surely, God loves .06 nearly as much as the .05," then we can conclude that God loves plots with confidence intervals as well.

In sum, adjusted confidence intervals computed using the *superb* framework are tailored to the testing situation. Consequently, they match a Welch test very closely in any situation for which an adjustment is known and applied. Thus, the golden rule is constant irrespective of the methodological details. It is therefore a universal rule of adjusted confidence intervals. This unique rule is in our opinion more useful than assessing proportions of overlap, as suggested by the numerous rules of thumb that many apply (and misapply).

We offer one final word. Goldstein and Healy (1995), Baguley (2012), and Franz and Loftus (2012) suggested dividing the interval width by 2. Tryon (2001) framed this as a human-factor necessity. The golden rule of half-width confidence intervals would be "If the extremity of one statistic's confidence interval does not overlap the extremity of another statistic's confidence interval, the two statistics should be interpreted as being statistically different" (also see Tryon, 2001, p. 375). However, first, the two rules are quite different and potentially confusing (one being based on inclusion, the other on overlap). Confusion can be avoided only by using a single type of interval in a consistent manner so that the reader can develop the correct automatisms (Cousineau & Larochelle, 2004). Second, half-width intervals are meaningful only when comparing a mean with other means. As we demonstrated in the examples, this excludes results that are compared with fixed values. Fixed values do not have confidence intervals, which precludes one touching the other. For these two reasons, we do not recommend half-width intervals.

Discussion

In this article, we focused solely on means and confidence intervals of the means. However, any functions can be provided. First, any function that, given a list of scores, returns a summary statistic can be used for the point estimates. That includes the median (and other measures of central tendency), but it could also be measures of spread (standard deviation, variance, or interquartile range), measures of skew (Fisher skew, Pearson skew), and so on. See the documentation for the full list of functions currently implemented. It can also be custom-made functions. As an example, see https://dcousin3.github.io/superb/articles/Vignette4.html for how the Cohen's d_1 statistic can be defined in R and this function's name given to the plotting function to generate point estimates. This summary statistic computes the standardized mean difference to a reference value (Goulet-Pelletier & Cousineau, 2018). Second, any measure of precision can also be used. It can be the standard error or the confidence interval (already available in the package) but could also be any custom-made function returning a length or an interval boundary. The above Web page shows how to define a confidence interval to d_1 (Goulet-Pelletier & Cousineau, 2018) and use it in plots. Note that the difference between two Cohens's d_1s is in fact the usual Cohen's d_p. It is therefore a convenient plot to illustrate standardized effect sizes.

When *superb* is used to show confidence intervals, the complementary framework is null hypothesis statistical testing. This specific use of *superb* could therefore be labeled a *frequentist superb* framework. However, the *superb* framework is also compatible with other inference methods. The *superb* framework is indifferent to which

inferential platform you choose. If you want to perform inference following null hypothesis significance testing, then confidence intervals are perfect. If you wish to remain neutral, you could choose to display standard errors. Both confidence intervals and standard errors are prepackaged in `superb` for more than a dozen descriptive statistics (including the mean, median, variance, skew, etc.). If you want to assess sampling error, opt for precision intervals (Cousineau & Goulet-Pelletier, 2021). Finally, for Bayesian examination, highest density intervals could be used, although they are not implemented at this time. Any custom-made function can be fed to `superb`.

Note that most estimators of precision included in the package at this time are parametric: They make assumptions regarding the distribution of the population. Alternatively, one could use nonparametric methods. One such method is the bootstrap method (Efron & Tibshirani, 1993). In a nutshell, this technique iteratively resamples data from the initial sample (with replacement) a large number of times (e.g., 10,000 times) to estimate the lower and upper boundaries containing the statistic of interest in 95% of the subsamples (Rousselet et al., 2019). Bootstrap intervals of the mean are nonparametric and require no assumptions other than the fact that the researcher has gathered an adequate sample. As we are reminded by Rousselet et al. (2019), bootstrap estimators are estimating the sampling distribution of the statistic, not the predictive distribution of the parameter. Hence, they are not confidence intervals but precision intervals (for the distinction, see Cousineau & Goulet-Pelletier, 2021).

In the package `superb`, this can be achieved by creating a custom-made interval function. Simple quantile bootstrap estimates are already packaged, but more elaborate algorithms (e.g., bias-corrected and accelerated) can be added by the user and provided to the plot function.

Keep in mind that summary statistics along with adequate error bars is the bare minimum. Informative plots should be supplemented with information illustrating the whole distribution whenever possible, as was done with jittered dots (Lane, 2019), rainclouds (Allen et al., 2019), and violins (Marmolejo-Ramos & Matsunaga, 2009). These layouts are packaged in `superb`, but custom layouts can be added. For example, the quantiles could be connected to illustrate the changes in every segment of the population (Rousselet et al., 2017). Other representations are possible, many of which are compatible with a summary plot containing error bars (e.g., Weissgerber et al., 2015).

Strengths and limitations of superb

The purpose of the `superb` package is to reunite in a single function all the known precision measures and all possible adjustment methods. Sadly, the *superb* framework is incomplete. For example, if sampling is stratified, an adjustment for stratified sampling should be applied. Unfortunately, we currently do not know how precise this method is relative to SRS. Likewise, when multilevel sampling is used (beyond two levels given that the two-level case is identical to CRS), there is no closed-form formula of precision at this time. These are two examples, but many more probably exist.

In addition, like any statistical method, precision estimation is based on assumptions. The confidence interval of the mean is based on a few assumptions. These assumptions could be tested, although some authors have warned against this practice (e.g., Delacre et al., 2018; Rochon et al., 2012, among others). The normality assumption can be assessed using a test of normality. When conditions from a within-subjects design are compared, the measurements must respect the sphericity assumption (methods CM and LM) or the compound symmetry assumption (method CA). These assumptions can be assessed with Mauchly's test of sphericity (Mauchly, 1940) or Winer's test of compound symmetry (Winer et al., 1991). The violation of these assumptions does not forbid the use of confidence intervals in graphs because they still provide a meaningful estimation of the variability around the displayed statistics. However, these confidence intervals should be interpreted with care.

As a last limitation, the *superb* framework is restricted to univariate (including repeated measures) variables. Multivariate studies in which interrelations between variables are important are not suited for *superb*. For example, there is no obvious way to plot a variance-covariance matrix with the present approach. Furthermore, it assumes that precision is also computed for a unique parameter. It is not obvious how the framework could represent parameter estimation or parameter precision from a hierarchical model with hyperparameters. A defining feature of the *superb* framework is that *superb* performs separate estimations. They can be pooled to make a Welch test, but they cannot interact.

The R package `superb` is one implementation of the *superb* framework. O'Brien and Cousineau (2014) provided a tool to implement difference-adjusted error bars and correlation-adjusted error bars within SPSS; Cousineau (2017) discussed a Mathematica implementation of *superb* that precluded the current R implementation. We hope to see other software adopt this framework.

Statistical tests can assess significance, but too often, they place the focus on whether the p threshold has been reached, which sometimes leads to using controversial manipulation, referred to as "p-hacking" (Simmons et al., 2011). Using visualization tools such as plots, researchers can better interpret the results, relying not only on statistical significance but also on the magnitude of the effect (Lakens, 2013; Loftus, 1993). The interpretation of

results should then be more balanced. A contrario, without measures of precision, interpretation is less nuanced (Fricker et al., 2019). However, researchers must always keep in mind that confidence intervals, like significance testing, may bring focus on null hypotheses, which rarely reflect the hypothesized effect (Beaulieu-Prévost, 2006). To be thorough, one must also consider alternative hypotheses, contextualizing the observed results under different models, and generate new predictions, a cornerstone of scientific discovery (Jamieson & Pexman, 2020).

By providing an easy-to-apply framework to adjust the error bars, we wish to ease the transition toward a better visualization and interpretation of summary statistics. Although the techniques presented here are accurate and based on formal mathematical demonstrations, very few statistical software offer anything but the stand-alone (and potentially contradictory) error bars. Science is difficult; there is no reason to make it more difficult by using stand-alone measures of precision whose interpretation is inconsistent with the methodology used. Despite its limitations, we believe that the *superb* framework is a major improvement over the use of stand-alone measures of precision supplemented with approximate heuristics. We elaborate powerful experimental methodologies to improve statistical power; it does not make sense that the measures of precision shown in plots hide these gains.

Appendix A: A Very Short Introduction to the *Superb* Library

To implement the *superb* framework, we devised an R library, *summary plots with adjusted error bar*, or in short, `superb`, available open source and open access from CRAN. See https://dcousin3.github.io/superb/ for complete documentation as well as examples.

To upload it, issue the command

```
install.packages("superb")
```

Once installed, only the following is necessary for subsequent sessions:

```
library(superb)
```

The library provides a main function, `superbPlot()`, which can be used to derive adjusted precision intervals in a wide variety of situations. It also contains a graphical user interface `superbShiny()`, which requires no programing experience; in the following, we present the former only.

Basic specification

The following is an example of the simplest use of `superbPlot`:

```
superbPlot (ToothGrowth,
  BSFactors = c("dose","supp"),
  variable = "len"
)
```

This command generates a mean plot with stand-alone 95% confidence intervals (by default) whereby the results can be compared with a prespecified value. It uses the `ToothGrowth` data set available in R, in which two between-subjects factors were manipulated (dose of vitamin C and type of supplement); the dependent variable is the tooth's length. The first six lines of data for this data set are:

```
head(ToothGrowth)
# len  supp   dose
# 1    4.2  VC 0.5
# 2   11.5  VC 0.5
# 3    7.3  VC 0.5
# 4    5.8  VC 0.5
# 5    6.4  VC 0.5
# 6   10.0  VC 0.5
```

Column names must match exactly those found in the data set and be given within quotes. Data must be in a data frame in wide format (i.e., if there are repeated measures, they must be in distinct columns, satisfying the *one-subject = one line* requirement of SPSS and SAS). For within-subjects designs, the factors must be entered with the number of levels, and all the measurement columns must be enumerated:

```
WSFactors = c("factor1(number of
  levels)", "factor2(number or
  levels)", etc.)
variables = c("column 1", "column 2",
  etc.)
```

Mixed designs can also be used with a mixture of `BSFactors` and `WSFactors`, but a maximum of four factors can be manipulated per plot.

Selecting the descriptive statistic and the type of error bars

It is possible to change the default descriptive statistics and the type of error bars with additional arguments (`statistic` and `errorbar`) :

```
superbPlot(ToothGrowth,
  BSFactors = c("dose", "supp"),
  variable = "len",
  statistic = "median",
  errorbar = "SE"
)
```

These specifications will display the median along with its respective standard error. The function names must be quoted and can be "mean", "median", "sd", and so on. Any function, including custom-made functions, can be used. Error bars can be "SE", "CI", or "bootstrapPI", but again, any interval function can be used.

Selecting the adjustments

The strength of superbPlot is that adjustments can be easily selected by providing a list of the desired adjustments, as in:

```
superbPlot(ToothGrowth,
  BSFactors = c("dose","supp"),
  variable = "len",
  adjustments = list( purpose =
    "difference")
)
```

This adjustment will generate difference-adjusted error bars (i.e., error bars that are 1.41 times longer) that can be used to compare groups. Other possible adjustments are purpose (single, difference or tryon), decorrelate (none, CM, LM or CA), samplingDesign (SRS or CRS), and popSize (the size of the population or Inf for infinite).

Adding directives to the plot

The plots returned by superbPlot are ggplot objects, and so it is possible to add additional directives to improve their appearance (Wickham, 2016). Here is an example showing 99.9% confidence intervals of the mean (set with the gamma argument) containing various directives (issue the command library(ggplot2) first):

```
superbPlot(ToothGrowth,
  BSFactors = c("dose","supp"),
  variables = "len",
  statistic = "mean",
  errorbar = "CI",
  gamma = .999,
  adjustments = list(purpose =
    "difference"),
  plotStyle = "raincloud",
  errorbarParams = list(width = .2,
    size = 1.5, colour = "gray"),
  pointParams = list(linetype = 3,
    colour = "black", size = .5)
) +
coord_cartesian( ylim = c(0, 40) ) +
xlab("Dose") + ylab("Tooth Growth") +
  labs(title = "Tooth Growth example")+
theme(axis.text.x = element_
  text(size=30, colour="red"))
```

The argument errorbarParams provides a list of ggplot directives to be sent to the display of the error bars only; the argument pointParams is the same for the display of the summary statistics. The remaining directives that follow the + after the command are general directives that set the vertical range and the axis labels and modify the plot's theme.

These are just a few of superb capabilities. Visit dcousin3.github.io/superb/ for the full documentations and more examples.

Appendix B: Four Steps to Obtain superb-Compliant Adjustments

Algorithm B.1 provides an overview of the steps involved to obtain a superb plot. These steps are described in the next sections.

1- Organize the data one $n \times J$ matrix per group	
Decide if you want to decorrelate the data	Yes: Equations B.1, B.2
Decide if you want to pool the estimates of precision	Yes: Equation B.3
2- For each group, obtain stand-alone precision lengths in a $1 \times J$ vector	Equations B.4a or B.4b
If needed, decide your confidence level (typically 95%)	
3- For each group, get adjustments to the stand-alone precision:	
a- Will comparisons be made to other sample means?	Equation B.5
b- Is correlation taken into account by a multiplicative adjustment?	Equation B.6
c- Are the groups obtained through cluster sampling?	Equation B.7
d- Is the population of finite size?	Equation B.8
4- Multiply the stand-alone lengths by all the adjustments	Equation B.9

Note: in Algorithm 1, n is the group size, and J is the number of repeated measure conditions. Step 3b implies that data were not decorrelated in Step 1.

Algorithm B.1: Steps to compute standard error of means and confidence intervals

Step 1: organize and decorrelate the data set. One convenient way to organize the data set is as a list of data matrices, one data matrix per group. Each data matrix has n lines, one per subject, and J columns, one per measurement in a repeated measures design. The J columns can be levels from a single or multiple factors. The groups can be of different sizes; for simplicity, we use n without subscript for the size of a given group.

In a repeated measures design, the measurements are often positively correlated such that participants who obtained high scores on one measurement tend to obtain high scores on other measurements as well. If measurements are to be compared with one another, it is thus possible to focus on within-subjects variance only by taking out between-subjects variances, which can be quantified from correlation.

One method advocated by Cousineau (2005; Morey, 2008; for reviews, also see Baguley, 2012; Bakeman & McArthur, 1996; Loftus & Masson, 1994) is to decorrelate the data set. Let $\mathbf{X}$ be a data matrix for one group, composed of J columns and n lines. The following transformation returns a data matrix $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{X} - E(\mathbf{X}) + E(E(\mathbf{X})), \tag{B.1}$$

in which $E(\mathbf{X})$ is a $1 \times n$ vector of means (one mean per subject) and $E(E(\mathbf{X}))$ is the grand mean. This transformation is a *subject-centering transformation* because it removes individual differences (Abdi, 2010). However, as was demonstrated in Morey (2008), the standard errors obtained from the $\mathbf{Y}$ matrix are biased downward, and thus, a second transformation can be used to rectify variability (for details see, Cousineau & O'Brien, 2014; for implementation in R, Matlab, Mathematica, and SPSS, see O'Brien & Cousineau, 2014). It is given by

$$\mathbf{Z} = \sqrt{\frac{J}{J-1}}(\mathbf{Y} - E(\mathbf{Y}')) + E(\mathbf{Y}'), \tag{B.2}$$

in which $E(\mathbf{Y}')$ is a $J \times 1$ vector of means (one per measurement). It is called a *bias-correction transformation*.

These two steps together were called the *Cousineau-Morey* method by Baguley (2012). These two steps are repeated for each group separately. This method is based on the assumption that the variance-covariance matrix is spherical.

Another method advocated by Loftus and Masson (1994) consists in first decorrelating the data (as per Equation B.1) and correcting the bias (as per Equation B.2), then pooling the standard deviation into a common standard deviation. This additional step can be done with a *pool-standard deviation transformation* given by

$$\mathbf{W} = \frac{S_p}{\mathbf{S}}(\mathbf{Z} - E(\mathbf{Z}')) + E(\mathbf{Z}'), \tag{B.3}$$

in which $\mathbf{S}$ is a $J \times 1$ vector containing all the standard deviations and S_p is the pooled standard deviation given by

$$S_p = \sqrt{\frac{1}{J(n-1)}\sum_{i=1}^{J}(n-1)S_i^2},$$

in which n is the number of participants, J is the number of measurements pooled, and S_i is the standard deviation of the ith measurements.

Cousineau (2005) suggested avoiding pooling the standard deviations, but Loftus and Masson (1994) recommended pooling within-subjects factors. Currently, there is no consensus, and both alternatives are considered legitimate.

Both Cousineau-Morey and Loftus-Masson methods are unable to illustrate whether the sphericity assumption holds. This assumption must be met when performing a test of means on repeated measures. Therefore, a measure of sphericity should be provided, such as the Greenhouse-Geiser or the Hyunh-Feldt epsilons (Huynh & Feldt, 1976; Lane, 2016). The function `superbPlot` issues notes that indicate ϵ when relevant.

Step 2: obtain stand-alone precision length. From the raw data matrix or the decorrelated data matrix, one can obtain the stand-alone precision lengths, one per column in the matrix. Precision can be given with standard error or confidence interval. This step is repeated over all groups.

The stand-alone standard error of the mean (noted SE_M) is given by

$$SE_M = \frac{S}{\sqrt{n}}, \tag{B.4a}$$

in which S is the sample standard deviation and n is the sample size. The stand-alone confidence interval of the mean is given by

$$CI_{length} = \frac{S}{\sqrt{n}} \times t_c, \tag{B.4b}$$

in which t_c is based on the degree of freedom. Note that when cluster randomized sampling is used, the degree

of freedom $n - 1$ should be replaced by the number of clusters $(k - 1)$.

For descriptive statistics other than the mean, Equations B.4a and B.4b are different (for a list, see Harding et al., 2014). When error bars are asymmetrical (e.g., the variance), lengths should be computed both below the statistic and above the statistic.

Step 3: compute the adjustments. The error bar lengths obtained in Step 2 are correct only in the stand-alone situation. In other situations, they must be adjusted using multiplicative terms. These adjusted precision measures are said to be *adjusted* to take into account the actual methodology of the experiment and the sampling and the size of the population.

a. Adjustment for purpose

One adjustment concerns the purpose of the precision intervals. Are they going to be used to examine a mean in isolation (or relative to a fixed value such as a criterion), or are they going to be used to compare means with other means?

When comparing a mean with another mean, both have imprecision. To combine this imprecision, the error variances must be adjusted to represent the error variance of a difference between two means. Assuming that the variances are homogeneous, this is done by multiplying one error variance by 2 or, equivalently, the standard error by $\sqrt{2}$. When the variances are heterogeneous, this can be done with Tryon's (2001) 2 *E* adjustment (*E* is multiplied by 2 because *E* was designed to return half-width intervals), in which

$$E = \frac{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}{\sqrt{\frac{S_1^2}{n_1}} + \sqrt{\frac{S_2^2}{n_2}}} = \frac{\sqrt{SE_1^2 + SE_2^2}}{\sqrt{SE_1^2} + \sqrt{SE_2^2}}.$$

Thus, the adjustment is

$$Adj_{purpose} \leftarrow \begin{cases} 1 & \text{if the mean is compared to a fixed value} \\ \sqrt{2} & \text{if the mean is to be compared to other means} \\ 2E & \text{... and if the variances are heterogeneous.} \end{cases} \quad \text{(B.5)}$$

Note that when the data are decorrelated (Step 1), information from other measurements is used to narrow the precision intervals. Thus, it indirectly implies that means will be compared with other means so that the difference adjustment should always be used in conjunction to the correlation adjustment.

b. Adjustment for repeated measures

In recent work (Cousineau, 2019), an alternative to decorrelation was proposed that does not imply transformations of the raw data. This method is consequently simpler. However, it requires that the data set satisfies the compound symmetry assumption, a more restrictive assumption than sphericity. It is referred to by the initials CA in `superbPlot`. If used, the data will not be transformed; instead, a multiplicative adjustment is returned:

$$Adj_{correlation} \leftarrow \begin{cases} 1 & \text{for between-subject designs or decorrelated data} \\ \sqrt{1-\bar{r}} & \text{for within-subject design with CA method,} \end{cases} \quad \text{(B.6)}$$

in which $\bar{r}$ is the average pairwise correlation across all the pairs of measurements in a given group.

If such adjustment is used, Step 1 must be ignored. In other words, the decorrelation of the data set and the CA adjustment are mutually exclusive adjustments.

c. Adjustment for sampling design

A common sampling design is *simple randomized sampling*, in which subjects are randomly selected from the population. There exist two major alternatives to simple randomized sampling: cluster sampling and stratified sampling.

Cluster sampling consists of picking clusters of subjects randomly rather than taking subjects randomly one by one. Cluster sampling is used, for instance, in educational studies in which classrooms are sampled rather than students. Nested designs and hierarchical analyses are often performed on data obtained from cluster sampling. The simplest form of cluster sampling is the *cluster randomized sampling* in which the clusters are randomly chosen but all the subjects of the sampled clusters are measured and analyzed. The influence of clusters can be measured by the intraclass correlation (or more aptly the intracluster correlation [ICC]). An ICC close to 1 (the maximum) indicates that participants within the clusters are similar and quite different from the participants in other clusters. Reviews suggest that ICC is commonly between 0.1 and 0.3 (see e.g., Goulet & Cousineau, 2019; Hedges & Hedberg, 2007).

Cousineau and Laurencelle (2016) derived the adequate correction to standard errors using the number of clusters *k*, the number of subjects per cluster *m*, and ICC. The correction factor, called $\lambda_{k,m}(\text{ICC})$, is given by

$$\lambda_{k,m}(\mathrm{ICC}) = \sqrt{\frac{1+(m-1)\,\mathrm{ICC}}{1-\left(\frac{m-1}{km-1}\right)\mathrm{ICC}}}.$$

Cousineau and Laurencelle (2016) indicated how ICCs can be obtained in R and in SPSS and the complete formula when clusters are of unequal sizes.

When the data are repeated measures, it is possible to assume the homogeneity of intraclass correlations. When it is assumed, the ICCs obtained from all the measurements are considered independent estimates of the same true ICC, and ICC(1, k) is used to improve estimation (Shrout & Fleiss, 1979).

Finally, stratified sampling is a method in which the sample matches the population on some indicator variable or variables. For example, the age of the participants could be used so that people of various ages are equally represented in the sample as they are in the general population. Unfortunately, there is no known formula to adjust precision in stratified sampling (Thompson, 2012).

In summary,

$$Adj_{sampling} \leftarrow \begin{cases} 1 & \text{for } \textit{simple randomized sampling} \\ \lambda_{k,m}(ICC) & \text{for } \textit{cluster randomized sampling} \\ ? & \text{for } \textit{stratified sampling}. \end{cases} \quad \text{(B.7)}$$

Unless ICC is below 0 (which is very unlikely), the correction factor for cluster randomized samples is always larger than 1 so that precision in cluster randomized sampling is weaker than in simple randomized sampling.

d. Adjustment for population size

If the sampled population has a finite size, it is possible that all members of the population have been sampled (if $n = N$, in which N is the size of the whole population). When this is the case, the precision of the mean is absolute because no information is missing. Thus, error bar lengths should tend to 0 as the proportion of sampled data from the population tends to 1 (Kish, 1965; Thompson, 2012).

To that end, use the following adjustment:

$$Adj_{popSize} \leftarrow \begin{cases} 1 & \text{for a population of infinite size} \\ \sqrt{1-n/N} & \text{for a population of finite size.} \end{cases} \quad \text{(B.8)}$$

Step 4: integrate all the adjustments. The only step remaining is to adjust the length of the error bars w with the adjustments found in Step 3:

$$w \leftarrow w \times Adj_{purpose} \times Adj_{correlation} \times Adj_{sampling} \times Adj_{popSize}. \quad \text{(B.9)}$$

When upper and lower lengths are different, both are adjusted in the same manner.

Transparency

Action Editor: Brent Donnellan
Editor: Daniel J. Simons
Author Contributions
D. Cousineau contributed the code and the first draft of the manuscript. M-A. Goulet and B. Harding tested the library and contributed to the text. All of the authors approved the final manuscript for submission.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
This research project was funded by M-A. Goulet's Ontario Graduate Scholarship and a Fonds de Recherche du Québec sur la Nature et les Technologie fellowships, B. Harding's Natural Sciences and Engineering Research Council fellowship, and D. Cousineau's Conseil pour la Recherche en Sciences Naturelles et en Génie Discovery Grant.
Open Practices
Open Data: not applicable
Open Materials: https://github.com/dcousin3/superb
Preregistration: not applicable
All materials have been made publicly available via GitHub and can be accessed at https://github.com/dcousin3/superb. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Denis Cousineau https://orcid.org/0000-0001-5908-0402
Marc-André Goulet https://orcid.org/0000-0001-9958-9182

Acknowledgments

We thank Rick Hoekstra, Daniel Lakens, and Guillaume Rousselet for their comments and Félix Chiasson for advice on ggplots.

Note

1. Compound symmetry is the assumption that variances are homogeneous across measurements and also that pairwise correlations are homogeneous across pairs of measurements. Sphericity, the assumption behind repeated measures analysis of variance, assumes that the variances of the differences between

pairs of measurements are homogeneous. Compound symmetric data are necessarily spherical, and situations with only two repeated measures are also necessarily spherical. Note that compound symmetry can be tested using the Winer test (Winer et al., 1991, p. 517; implemented in the R package).

References

Abdi, H. (2010). The Greenhouse-Geisser correction. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 1–10). SAGE.

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. (2019). *RainCloudPlots tutorials and codebase* (Version v1.1). Zenodo. http://doi.org/10.5281/zenodo.3368186

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*, 158–175. https://doi.org/10.3758/s13428-011-0123-7

Bakeman, R., & McArthur, D. (1996). Picturing repeated measures: Comments on Loftus, Morrison and others. *Behavior Research Methods, Instruments, & Computers*, *28*, 584–589. https://doi.org/10.3758/BF03200546

Beaulieu-Prévost, D. (2006). Confidence intervals: From tests of statistical significance to confidence intervals, range hypotheses and substantial effects. *Tutorials in Quantitative Methods for Psychology*, *2*, 11–19.

Bryk, A., & Raudenbush, S. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods*. SAGE.

Cochran, W. G. (1953). *Sampling techniques*. John Wiley & Sons.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–45. https://doi.org/10.20982/tqmp.01.1.p042

Cousineau, D. (2017). Varieties of confidence intervals. *Advances in Cognitive Psychology*, *13*, 140–155. https://doi.org/10.5709/acp-0214-z

Cousineau, D. (2019). Correlation-adjusted standard errors and confidence intervals for within-subject designs: A simple multiplicative approach. *The Quantitative Methods for Psychology*, *15*(3), 226–241. https://doi.org/10.20982/tqmp.15.3.p226

Cousineau, D. (2020). How many decimals? Rounding descriptive and inferential statistics based on measurement precision. *Journal of Mathematical Psychology*, *97*, Article 102362. https://doi.org/10.1016/j.jmp.2020.102362

Cousineau, D., & Goulet-Pelletier, J.-C. (2021). A study of confidence intervals for Cohen's dp in within-subject designs with new proposals. *The Quantitative Methods for Psychology*, *17*, 51–75. https://doi.org/10.20982/tqmp.17.1.p051

Cousineau, D., & Larochelle, S. (2004). Visual-memory search: An integrative perspective. *Psychological Research*, *69*, 77–105. https://doi.org/10.1007/s00426-003-0170-5

Cousineau, D., & Laurencelle, L. (2016). A correction factor for the impact of cluster randomized sampling and its applications. *Psychological Methods*, *21*, 121–135. https://doi.org/10.1037/met0000055

Cousineau, D., & O'Brien, F. (2014). Error bars in within-subject designs: A comment on Baguley (2012). *Behavior Research Methods*, *46*, 1149–1159. https://doi.org/10.3758/s13428-013-0441-z

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. https://doi.org/10.1177/0956797613504966

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Journal of Psychology*, *217*, 15–26. https://doi.org/10.1027/0044-3409.217.1.15

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–574. https://doi.org/10.1177/00131640121971374

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*, 170–180. https://doi.org/10.1037/0003-066X.60.2.170

Delacre, M., Lakens, D., & Leys, C. (2018). Why psychologists should by default use Welch's t-test instead of the Student's t-test. *International Review of Social Psychology*, *30*, 92–101. https://doi.org/10.5334/irsp.82

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods for Psychology*, *12*, 30–38. https://doi.org/10.20982/tqmp.12.1.p030

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review*, *4*, 330–341. https://doi.org/10.3758/BF03210790

Franz, V. H., & Loftus, G. R. (2012). Standard errors and confidence intervals in within-subjects designs: Generalizing Loftus and Masson (1994) and avoiding the biases of alternative accounts. *Psychonomic Bulletin & Review*, *19*, 395–404. https://doi.org/10.3758/s13423-012-0230-1

Fricker, R. D., Jr., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the statistical analyses used in *Basic and Applied Social Psychology* after their *p*-value ban. *The American Statistician*, *73*, 374–384. https://doi.org/10.1080/00031305.2018.1537892

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society A*, *158*, 175–177. https://doi.org/10.2307/2983411

Goulet, M-A., & Cousineau, D. (2019). The power of replicated measures to increase statistical power. *Advances in Methods and Practices in Psychological Sciences*, *2*(3), 199–213. https://doi.org/10.1177/2515245919849434

Goulet-Pelletier, J-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen's d family. *The Quantitative Methods for Psychology*, *14*, 242–265. https://doi.org/10.20982/tqmp.14.4.p242

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337–350. https://doi.org/10.1007/s10654-016-0149-3

Harding, B., Tremblay, C., & Cousineau, D. (2014). Standard errors: A review and evaluation of standard error estimators using Monte Carlo simulations. *The Quantitative Methods*

for Psychology, *10*, 107–123. https://doi.org/10.20982/tqmp.10.2.p107

Harding, B., Tremblay, C., & Cousineau, D. (2015). The standard error of the Pearson skew. *The Quantitative Methods for Psychology*, *11*, 32–37. https://doi.org/10.20982/tqmp.11.1.p032

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87. https://doi.org/10.3102/0162373707299706

Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-splot designs. *Journal of Educational Statistics*, *1*, 69–82. https://doi.org/10.3102/10769986001001069

Jamieson, R. K., & Pexman, P. M. (2020). Moving beyond 20 questions: We (still) need stronger psychological theory. *Canadian Journal of Experimental Psychology*, *61*, 273–280. https://doi.org/10.1037/cap0000223

Kish, L. (1965). *Survey sampling*. John Wiley & Sons.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, Article 863. https://doi.org/10.3389/fpsyg.2013.00863

Lane, D. M. (2016). The assumption of sphericity in repeated-measures designs: What it means and what to do when it is violated. *The Quantitative Methods for Psychology*, *12*, 114–122. https://doi.org/10.20982/tqmp.12.2.p114

Lane, D. M. (2019). Graphing within-subjects effects. *The Quantitative Methods for Psychology*, *15*, 174–187. https://doi.org/10.20982/tqmp.15.3.p174

Loftus, G. R. (1993). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers*, *25*, 250–256. https://doi.org/10.3758/BF03204506

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490. https://doi.org/10.3758/BF03210951

Marmolejo-Ramos, F., & Matsunaga, M. (2009). Getting the most from your curves: Exploring and reporting data using informative graphical techniques. *Tutorials in Quantitative Methods for Psychology*, *5*, 40–50.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, *11*, 200–209.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*, 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Nathoo, F. S., Kilshaw, R. E., & Masson, M. E. J. (2018). A better (Bayesian) interval estimate for within-subject designs. *Journal of Mathematical Psychology*, *86*, 1–9. https://doi.org/10.1016/j.jmp.2018.07.005

O'Brien, F., & Cousineau, D. (2014). Representing error bars in within-subject designs in typical software packages. *The Quantitative Methods for Psychology*, *10*, 56–67. https://doi.org/10.20982/tqmp.10.1.p056

Pfister, R., & Janczyk, M. (2013). Confidence intervals for two sample means: Calculation, interpretation, and a few simple rules. *Advances in Cognitive Psychology*, *9*, 74–80. https://doi.org/10.2478/v10053-008-0133-x

Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, *12*, Article 81. https://doi.org/10.1186/1471-2288-12-81

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284. https://doi.org/10.1037/0003-066X.44.10.1276

Rouanet, H., & Lecoutre, B. (1983). Specific inference in ANOVA: From significance tests to Bayesian procedures. *British Journal of Mathematical and Statistical Psychology*, *36*, 252–268.

Rousselet, G. A., Penet, C. R., & Wilcox, R. R. (2017). Beyond differences in means: Robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience*, *46*, 1738–1748. https://doi.org/10.1111/ejn.13610

Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2019). *A practical introduction to the bootstrap: A versatile method to make inferences by using data-driven simulations*. PsyArXiv. https://doi.org/10.31234/osf.io/h8ft7

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Thompson, S. K. (2012). *Sampling* (3rd ed.). John Wiley & Sons.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferencial confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386. https://doi.org/10.1037/1082-989X.6.4.371

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar on line graphs: Time for a new data presentation paradigm. *PLOS Biology*, *13*, Article e10021128. https://doi.org/10.1371/journal.pbio.10021128

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350–362. https://doi.org/10.2307/2332010

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. https://doi.org/10.1037/h0027060

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design*. McGraw-Hill.

Yang, B. W., Vargas Restrepo, C., Stanley, M. L., & Marsh, E. J. (2021). Truncating bar graphs persistenly misleads viewers. *Journal of Applied Research in Memory and Cognition*. Advance publication. https://doi.org/10.1016/j.jarmac.2020.10.002