University of North Georgia
Nighthawks Open Institutional Repository

Faculty Publications | Department of Biology

Winter 2-7-2020

Making the Error Bar Overlap Myth a Reality: Comparative Confidence Intervals

Frank Corotto
University of North Georgia, frank.corotto@ung.edu

Follow this and additional works at: https://digitalcommons.northgeorgia.edu/bio_facpub

 Part of the Biology Commons, Social and Behavioral Sciences Commons, and the Statistics and Probability Commons

Recommended Citation

Corotto, Frank, "Making the Error Bar Overlap Myth a Reality: Comparative Confidence Intervals" (2020). *Faculty Publications*. 1.
https://digitalcommons.northgeorgia.edu/bio_facpub/1

This Article is brought to you for free and open access by the Department of Biology at Nighthawks Open Institutional Repository. It has been accepted for inclusion in Faculty Publications by an authorized administrator of Nighthawks Open Institutional Repository.

Making the Error Bar Overlap Myth a Reality:
Comparative Confidence Intervals

Frank S. Corotto
University of North Georgia,
Dahlonega, Georgia, 30597
Frank.Corotto@ung.edu

Introduction

Error bars are often misinterpreted (Belia et al. 2005; Cumming et al. 2004). A common myth is that when error bars for two samples do not overlap, the difference is statistically meaningful, a term I use in place of *statistically significant*. This overlap rule is really an overlap myth; the rule does not hold true for any type of conventional error bar. There are rules of thumb for estimating *P* values (Cumming et al. 2007), but it would be better to show bars for which that overlap rule holds true. We could quickly assess the statistical meaningfulness of a pattern.

If we want the overlap rule to hold true, what should we plot as error bars? John Tukey gave the answer (see Benjamini and Braun 2002) and suggested that *interference notches* would be a good way to show the intervals graphically (Tukey 1993). Others unknowingly repeated Tukey's work in different ways (Austin and Hux 2002; Knoll et al. 2011) with Schuun (1999) using *statistical significance bars* and Tryon (2001) *inferential confidence intervals* in place of Tukey's notches.

None of the proposed terms for these error bars is ideal. All confidence intervals are inferential, statistical significance is widely misunderstood (which is why I use *meaningfulness* instead)[1], and Tukey's notches cannot be created with spreadsheets. I propose *comparative confidence intervals* (CCIs), preceded by α, as in *0.05 CCIs.* The use of α reminds us that CCIs are not conventional confidence intervals.

To facilitate the broader use of comparative confidence intervals, I show here how to calculate the CCIs, how the intervals can be used in a variety of settings, and how they can be validated. I also explain why box-and-whiskers plots are good way to show CCIs, in place of Tukey's notches. Schunn (1999) touched on some of the topics I cover here, but his approach was mathematical. To make a better case for comparative confidence intervals, I use figures instead.

Conventional Confidence Intervals

To understand how comparative confidence intervals are calculated, we first have to understand conventional confidence intervals. Conventional intervals are calculated by performing null hypothesis tests backwards, often single-sample *t* tests. The formula for *t* is as follows.

[1]I use *meaningfully different*, *statistically meaningful*, and *statistically different.* "We in the behavioral sciences should 'give' this word [significant] back to the general public." R. Kline. 2004. Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research. American Psychological Association, 325 pp. See p. 87. Kline would use "statistical difference", but it is awkward to turn that around and say that a difference is statistical.

$$t = \frac{|\bar{x} - \mu|}{SE}$$

The outcome of a study is the sample mean ($\bar{x}$), the prediction based on the null hypothesis is the hypothetical population mean (μ), and *SE* represents standard error. To perform a *t* test, we start with the difference between our outcome and our prediction, e.g., the numerator in the equation. We divide that by standard error and use the resulting *t* value to find *P*. To calculate a confidence interval, we start with a *P* value (α), find the *t* value that goes with that *P* value, i.e., the *critical value* of *t*, and multiply it by standard error. Since standard error is in the denominator of the formula for *t*, when we multiply the critical value of *t* by standard error, standard error cancels out. We are left with the numerator in the formula, which is half of the confidence interval. The mean would be shown plus and minus that half-interval. What does that half-interval show? It shows the numerator in the formula for *t* that corresponds to $P = \alpha$, i.e., the smallest difference between our prediction and our outcome that would yield a finding of $P \leq \alpha$.[2] The full interval contains the range of differences that would lead to a finding of statistical meaningfulness, and we calculated the half-interval by performing a *t* test backwards.

Why would that range of differences be important? If we set α to 0.05, over a lifetime of constructing 95% confidence intervals around sample means, the population means will be outside of those intervals 5% of the time. Similarly, if we set α to 0.05, over a lifetime of testing true null hypotheses, we will incorrectly reject 5% of them. Confidence intervals and *P* values both show the results of null hypothesis tests.

Misconceptions pertaining to confidence intervals parallel those pertaining to *P*. Neither one gives us a probability of anything on any one occasion. A population mean is either within a confidence interval or it is not, the same way we either make a type I error or we do not. With both confidence intervals and *P* values, the results are either entirely due to sampling error or they are not. The probability in all of these cases is either one or zero. As bad as these misconceptions are statements like *We can be 95% confident that . . .* What does it mean to be 95% confident of something? It means *You won't understand, so I will tell you something meaningless instead.*

Simultaneous Confidence Intervals

Comparative confidence intervals would be most useful when there are multiple comparisons being made. We could easily assess the statistical meaningfulness of a pattern. When there are multiple comparisons, however, we cannot base our confidence intervals on *t* tests. To explain, suppose we collect samples A, B, and C and compare A with B, A with C, and B with C by performing three *t* tests. The cumulative or *familywise* error rate would be 0.14, not 0.05 (for why

[2] Except in the following paragraph, I do not speak of rejecting or failing to reject the null because many null hypotheses cannot be true. There is a large body of literature on that topic. A commonly cited, early source is P. Meehl. 1967. Theory-testing in psychology and physics: a methodological paradox. Philosophy of Science, 34: 151–159.

it is not 3 × 0.05 = 0.15, see Zar 2010, pp. 189,190). To keep familywise error at α, instead of performing *t* tests backwards to get our intervals, we can perform multiple comparisons tests backwards. Good multiple comparisons tests hold familywise error at α. The result would be *simultaneous* confidence intervals, simultaneous in that they have been corrected for multiple comparisons. Here I use Tukey-Kramer tests, because the Tukey test is highly regarded (Zar 2010, p. 232), and the Tukey-Kramer variation allows sample size to vary.

To illustrate the calculation of simultaneous confidence intervals, I created eight samples with similar variances but different sample sizes and performed a 1-way ANOVA (Appendixes A and B). The denominator in the resulting *F* ratio is variously termed *mean square error, MS error*, or simply *the error term.* The error term is important later, but for now we need the degrees of freedom associated with it, which is 52 (Appendix B). We use those 52 degrees of freedom; the number of categories compared by the ANOVA, which is eight (typically shown as *k* in tables); and α (we will use 0.05); to find the corresponding critical value of *q* (q_{CV}; use the table of critical values of *q* not *t*). In this case that critical value is 4.466. We calculate standard error with the Tukey-Kramer formula, which follows.

$$SE = \sqrt{\left(\frac{MS\ error}{2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

The two sample sizes are indicated by n_1 and n_2, and *MS error* is the denominator in the *F* ratio (Appendix B). We will use sample A and sample F (Appendix A) as an example.

$$SE = \sqrt{\left(\frac{2.308}{2}\right)\left(\frac{1}{5} + \frac{1}{7}\right)}$$

$$SE = 0.629$$

Standard error multiplied by q_{CV} yields a half simultaneous confidence interval of 2.809.

For both Tukey and Tukey-Kramer tests, the test statistic *q* is calculated with the formula below

$$q = \frac{|\bar{x}_A - \bar{x}_B|}{SE}$$

in which $\bar{x}_A$ and $\bar{x}_B$ are the two sample means. By performing a Tukey-Kramer test backwards, we have solved for the numerator in the formula for *q* that corresponds to *P* = 0.05. The difference between the two sample means ($\bar{x}_A - \bar{x}_F = -3$) plus and minus the half simultaneous confidence interval (2.809) contains the range of differences between the two sample means that would have led to an outcome of statistical meaningfulness (–5.809 through –0.191).

Figure 1 illustrates a common way to show the results. The differences between every pair of sample means are plotted along with a family of simultaneous confidence intervals based on different α's. For samples A and F, –3 is plotted along with bars that end at –5.809 and –0.191,

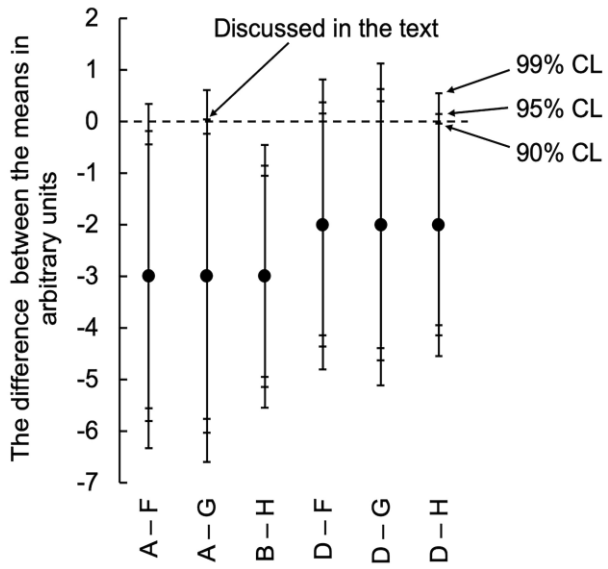

Figure 1. Some of the pairwise differences among the sample means in Appendix A, along with conventional simultaneous confidence intervals. CL = confidence limit.

the 95% simultaneous confidence limits. The fact that zero lies outside of the 95% simultaneous confidence interval but inside the 99% interval shows that *P* is less than 0.05 but greater than 0.01, respectively. The actual *P* value is 0.028 (Appendix D). The error bars illustrate the results of Tukey-Kramer tests.

Comparative Confidence Intervals

One problem with plots like the one in Figure 1 is that we must think about what is being subtracted from what to interpret the signs of the outcomes. It is sample mean A minus sample mean F, so the negative difference means that F is greater than A, and not the other way around. Another problem is that, by showing the differences between the means, we cannot compare the means themselves by eye; larger patterns are obscured. It would be better to plot means with comparative confidence intervals. To calculate CCIs, we simply divide half simultaneous confidence intervals by two. Here is why.

Consider the comparison of samples A and G in Figure 1. The difference between the means is –3 and the upper 95% simultaneous confidence limit lies almost on zero. Suppose that limit were exactly zero, i.e., $P = 0.05$, and the means themselves were plotted rather than the difference between them. Those means would be separated by 3. If we want bars for which separation indicates that $P < 0.05$, how long should they? They should be half the length of the bar extending from –3 to zero. To calculate comparative confidence intervals, we calculate half simultaneous confidence intervals, and divide by two. Then we plot the CCIs around means, not differences. The idea goes back to John Tukey. Benjamini and Braun (2002) describe his thoughts as follows.

> *If there exists a distance beyond which the two means are considered separated, then an effective graphical display involves drawing an allowance equal to plus or minus half that distance around the mean, and noting whether the allowances of the pair of means being compared overlap.*

In the case of samples A and G, the "distance beyond which the two means are considered separated" is the half simultaneous confidence interval of 3. See also Figures 7 and 8 in Wainer (1996).

We will use sample A to show how CCIs are calculated. Because there is only one sample, we calculate standard error with the Tukey test's formula, which is as follows.

$$SE = \sqrt{\frac{MS\ error}{n}}$$

Here is the calculation for sample A.

$$SE = \sqrt{\frac{2.308}{5}}$$

$$SE = 0.679$$

If $\alpha = 0.05$, and there are eight groups, the critical value of q is 4.466, as we saw earlier That critical value multiplied by standard error yields a half simultaneous confidence interval of 3.032. We divide that half-interval by two and get 1.516. Bars that long would be plotted around sample A's mean of 2 to show 0.05 CCIs.

How to Plot Comparative Confidence Intervals

When using comparative confidence intervals, we must assess the degree to which error bars overlap with other error bars. This can be difficult if families of CCIs are plotted that correspond to different α's. One way to improve the situation is to plot just two intervals. I suggest 0.05 and 0.15 CCIs.[3] I chose 0.05 because it is a traditional α and 0.15 because it allows us to see close calls. If the 0.05 CCIs overlap, but the 0.15 CCIs do not, *P* is between 0.05 and 0.15; it may be worth increasing the sample size or conducting another study to further investigate. I prefer 0.15 to 0.10 because the 0.10 CCIs come too close to those for 0.05.

Another way to add clarity is to use box-and-whiskers plots. The boxes would show the 0.15 intervals and the whiskers 0.05. For example, if we compare sample A with sample E, the whiskers do not overlap, and $P = 0.022$ (Figure 2; Appendix C). If we compare samples D and H, the whiskers overlap, but the boxes do not, and $P = 0.085$.

Differing Sample Sizes

When solving for *P*, the Tukey-Kramer formula is used to calculate standard error. This means standard error may vary depending on which sample is being compared to which. For example, in the comparison of sample A with sample F, we found that standard error was 0.629. For samples

[3] The critical values for 0.15 are available at the bottom of this page https://digitalcommons.northgeorgia.edu/bio_facpub/1/

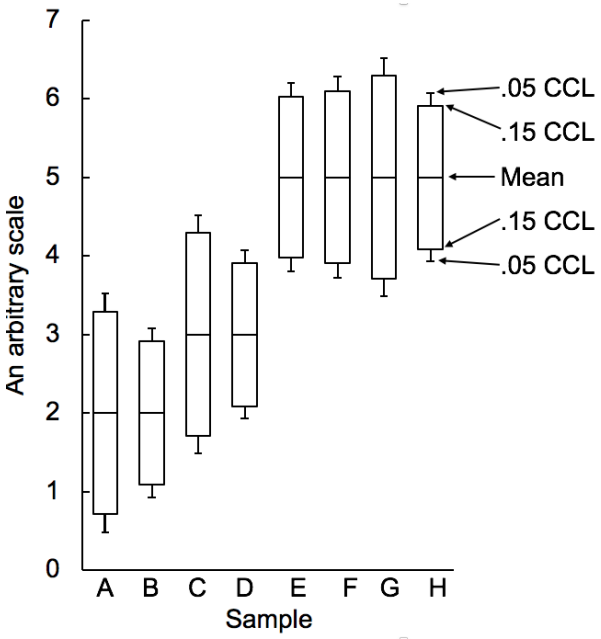

Figure 2. The means in Appendix A, along with comparative confidence intervals. CCL = comparative confidence limit; *n* = sample size.

A and G, however, standard error is 0.679. Consequently, the *conventional* intervals for the comparison of sample A with F are smaller than those for the comparison of A with G (Figure 1). To calculate the comparative confidence intervals in Figure 2, I used the Tukey test's formula to calculate standard error, with *n* varying according to each sample. Do the resulting intervals reflect the *P* values that would be obtained by performing Tukey-Kramer tests in which standard error varies? For example, can the CCIs for sample A be compared to those for both F and G when Tukey-Kramer tests for those comparisons would entail the use of different standard errors? To find out, I plotted the *P* values obtained from some of the Tukey-Kramer tests (Appendix D) as a function of the distance between the 0.05 and 0.15 CCIs, choosing results for which sample size varies and *P* values are low. The CCIs reflect the *P* values almost perfectly. Although sample A, for example, was compared to four different samples with four different sample sizes, the data for those comparisons line up with the rest (see the asterisks in Figure 3), and the curves as a whole

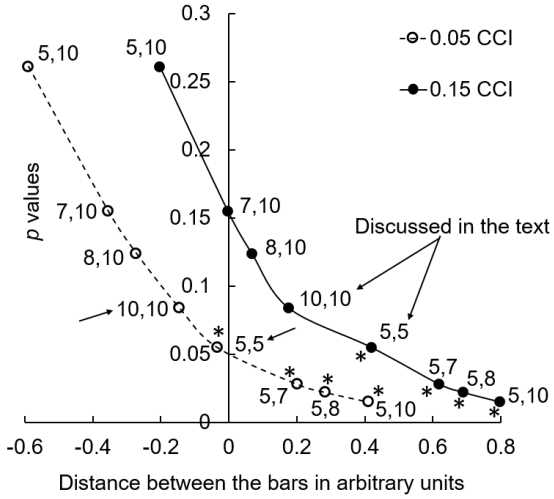

Figure 3. The distance between bars representing comparative confidence intervals reflects *P* values calculated with the Tukey-Kramer method. Negative *x*-values represent overlap of the bars. Asterisks indicate comparisons of sample A with other samples. Comparative confidence intervals are shown in Figure 2. Numbers indicate sample sizes for each pair of samples being compared. CCI: comparative confidence interval.

have *y*-intercepts of 0.05 and 0.15. Large differences in sample size create slight anomalies. In the case where both samples sizes are 10, the two curves come close to each other while, when both sample sizes are five, they are farther apart (indicated by arrows in Figure 3). When the size of each sample is used to calculate standard error, CCIs reflect the results of Tukey-Kramer tests well enough to serve their function, even when there are large differences in sample size.

Other Tests for Other Situations

Because comparative confidence intervals can be obtained by performing Tukey tests backwards, any way we perform null hypothesis tests forward, to obtain *P*, we can perform backwards, to get CCIs. If we were to compare a number of samples to a single reference (such as a control) and not to each other, we could calculate the intervals by performing Dunnett's test backwards. Dunnett's test is more powerful than Tukey's. What is important is that the critical values are based on the degrees of freedom associated with mean square error and the total number of groups, since those are the values that would be used to conduct Dunnett's tests in the forward direction, and standard error should be based on mean square error as well. If there are only two samples being compared, comparative confidence intervals are not too important—there is no larger pattern to assess—but we could perform two-sample *t* tests backwards to get comparative confidence intervals. It would just be a matter of using total degrees of freedom to find the critical value of *t* and using pooled variance to calculate standard error, since pooled variance is the equivalent of mean square error.

If sample size varies, standard error can be calculated for each sample based on each sample's size, i.e., as was done for the intervals shown in Figure 2. This method works just as well for intervals based on Dunnett's test and two-sample *t* tests as for intervals based on Tukey tests (the outcomes are similar to what is shown in Figure 3). Note that for both tests mean square error must be multiplied by two when calculating standard error. This is not the case when calculating standard error to obtain Tukey-based comparative confidence intervals.

Main Effects and Interactions

Sometimes null hypothesis tests only tell us what is already obvious once we plot our data. Where these tests are particularly helpful, however, is when there are multiple independent variables, i.e., a factorial design. Independent variables can have effects on their own, *main effects*, and they can affect each other; they can *interact*. It is often hard to judge by eye whether such an interaction is statistically meaningful or created by sampling error. We need to calculate *P*. To illustrate how we can use CCIs to show these *P* values, imagine we are testing three brands of tire, at the front and rear positions, and determining their longevity. If every possible combination of independent variables is represented, we have a factorial design (Figure 4).

With a factorial design, the averages for each combination of every independent variable are referred to as cell means, because they occupy cells in the matrix that illustrates the factorial design, e.g., brand A went an average of 40,000 miles in the front position (Figure 4). If we pool the data across the rows or columns, we can calculate *marginal* means that illustrate the main

		Brand of car tire			
		A	B	C	Marginal means and sample sizes
Position	Front	$\bar{x}$ = 40k n = 20	$\bar{x}$ = 50k n = 20	$\bar{x}$ = 50k n = 20	$\bar{x}$ = 46.7k n = 60
	Rear	$\bar{x}$ = 35k n = 19	$\bar{x}$ = 40k n = 20	$\bar{x}$ = 45k n = 20	$\bar{x}$ = 40k n = 59
	Marginal means and sample sizes	$\bar{x}$ = 37.5k n = 39	$\bar{x}$ = 45k n = 40	$\bar{x}$ = 47.5k n = 40	

Figure 4. A factorial design in which the longevity of three brands of car tire are compared at the front and rear positions. Longevity is in thousands (k) of miles. Sample size is indicated by *n*.

effects of each independent variable. For example, the average longevity of brand A is the average of its two cell means, 35,000 and 40,000 miles, or 37,500 miles, shown in the bottom margin in Figure 4. Similarly, we can pool sample sizes and illustrate them in the margins too. Understanding cell and marginal means and sample sizes will help us understand how to use comparative confidence intervals when there is a factorial design.

In the case of tire brand and position, we would analyze the results with a 2-way ANOVA, because there are two independent variables. The ANOVA would generate *F* ratios and *P* values for both of the main effects (*tire* and *position*) and also for the interaction. If there is a statistically meaningful main effect of *tire,* we might plot the marginal means of the three brands along with comparative confidence intervals to illustrate which brand is statistically different from which. The CCIs would be based on whatever multiple comparisons test we would use to compare the three brands. Here Tukey-Kramer tests would be appropriate because sample size varies. We would use the number of groups being compared (three) and the degrees of freedom associated with the error term for the main effect of *tire* to find the critical value of *q*. To calculate standard error, we would use the error term for the main effect of *tire* for variance; we are assuming equal variances, so the best estimate is that error term; and the marginal samples sizes for each group, e.g., 39, 40, and 40 in the example shown in Figure 4. The resulting CCIs would illustrate the results of Tukey-Kramer tests used to investigate the main effect of *tire*.

There would be no reason to investigate the main effect of position, since our interest is in tire brand, but if we did want to plot the marginal means for *front* and *rear*, we could base our comparative confidence intervals on a two-sample *t* test in which the marginal means of 40,000 and 46,700 miles (Figure 4) serve as the sample means. To calculate standard error, we would use the marginal sample sizes of 59 and 60 and the error term for the main effect of *position* for variance. If there were three positions, as would be the case when towing a small trailer, then we would base our CCIs on Tukey-Kramer tests, not *t* tests.

If the interaction is statistically meaningful, it means that we can exclude chance as the sole cause of a difference between differences. For example, there is a greater difference between brands A and B when they are at the front position than when they are at the rear. Is that difference in differences statistically meaningful? Is that why $P \leq \alpha$ for the interaction? To find out, we might perform two sets of multiple comparisons tests, one for *front* and one for *rear*; or three two-sample *t* tests, one for each brand. To illustrate the results, we would simply base our CCIs on which of those two options we take. (It is not justifiable to do both.). Error bars must always be explained, so we would make clear that the error bars can only be used for comparing across the brands within each position or vise versa.

Areas for Future Study

I know of two situations in which there are problems with comparative confidence intervals. One is when there is heterogeneity of variance. Although Tukey tests are highly regarded (Zar 2010, p. 232), they are not robust when variance differs among samples, especially when sample size varies as well (Zar 2010, pp. 230, 231). The other problem area is when there are repeated measures, i.e., paired data[4], blocks, and other situations in which subjects are compared to themselves. A repeated measures ANOVA removes the variation among subjects from the analysis. This reduces the error term and increases power. The problem with repeated measures, though, is the requirement for *sphericity*: all samples must correlate to each other to the same degree. "Violation of [this assumption] is, unfortunately, common . . ." (Zar 1010 p. 274). Concerns regarding both sphericity and variance are often addressed the same way, by testing the null hypothesis that sphericity holds or that variances are uniform. This practice presents two problems. Because a null hypothesis is infinitely precise, many nulls cannot be correct,[5] so failing to reject them or accepting them makes no sense. More importantly, if we "fail to reject" and decide there is no problem with sphericity, we are asking if a difference is large enough to be important. A null hypothesis test cannot answer that question. See also O'Brien and Kaiser (1985, pp. 318, 331).

Varying Variance

Tukey tests work well when samples sizes are the same and variances are at least "similar" (Zar 2010, pp. 230, 231). We can plot Tukey-based comparative confidence intervals. If sample size varies and variances are not "similar", one option is to transform the data to achieve homogenous variances. We could plot means and CCIs of the transformed data. Transformation can change the nature of the question, however. For example, an interaction following a log transformation indicates proportional differences, rather than absolute ones. Transformation can also fail equalize variances. In the case of a rank transformation, the more the samples overlap, the less successful the transformation.

---

[4] Sphericity always holds when there are only two samples, total, being compared, but CCIs are not valuable in that circumstance.

[5] A commonly cited, early source is P. Meehl. 1967. Philosophy of Science, 34: 151–159

Another strategy begins with the Games and Howe nonparametric alternative to the Tukey test, to get *P* values, but what should we plot as error bars? Because the Games and Howe procedure begins with a rank transformation, not within each sample but across all the data pooled, we could plot means and comparative confidence intervals of those ranks. We would be plotting the results of Tukey tests on ranks, though, not Games and Howell tests. The validation strategy in Figure 3 would be necessary.

Repeated Measures

Repeated measures presents a more vexing problem than differences in variance—there is no transformation to correct for a lack of sphericity. The severity of the problem can be gauged with the Greenhouse-Geisser method, the Huynh-Feldt method, and others. Those methods produce a statistic, ε, which ranges from zero to one, with one indicating perfect sphericity. I know of no rule of thumb for deciding if ε is small enough to justify taking action. It is common to correct the ANOVA by multiplying both of the *F* ratio's degrees of freedom by ε. The more severe the problem, the lower the value of ε, and the greater the correction. One method that *might* work to create CCIs would be to multiply degrees of freedom by ε when finding the critical value of *q*. The CCIs would be corrected just like the ANOVA. Unfortunately, I have not seen this method in the literature, and I lack the expertise to test it.

Suppose we abandon comparative confidence intervals and show conventional intervals instead. We would use the error term for the repeated measure to calculate standard error (Loftus and Masson 1994) but the problem with sphericity remains. Franz and Loftus (2012) show how to calculate intervals that would be plotted around differences, i.e. as in Figure 1, with those intervals correcting for a lack of sphericity.

Summary

Much of what I have discussed here has been described before. From what I can tell, the strategies for addressing different sample sizes and the problems with Tukey tests are mine, as is my advocacy for box-and-whiskers plots and my suggestion of the validation strategy illustrated in Figure 3. Because comparative confidence intervals are calculated by performing null hypothesis tests backwards, the intervals have the potential to be based on tests other than those I discussed. When basing CCIs on other types of tests, the intervals can be validated with the analysis illustrated in Figure 3.

Confidence intervals are "the best reporting strategy" according to the American Psychological Association (APA 2010, p. 34). Conventional intervals that flank sample means provide a range of likely values for population means. When samples are compared, however, the relations among the sample means can be more important that the means themselves. When means are being compared to each other, comparative confidence intervals should be plotted, along with or instead of conventional intervals.

Null hypothesis testing has been debated for decades. In fields in which it is termed null hypothesis *significance* testing, always initialized to NHST, the practice of making thoughtless yes-or-no decisions based on *P* values was once rampant. With comparative confidence intervals, we can practice thoughtless NHST. We can break out the T-square and see what overlaps with what. At the other end of the spectrum, Loftus (1993) encouraged plotting means with standard error and abandoning null hypothesis tests. With CCIs, we can take Loftus's advice to an extreme. We can take in the big picture and never think about *P* values. Most of us will choose some strategy in between NHST and Loftus's, and CCIs will serve us well. Comparative confidence intervals make the APA's "best reporting strategy" even better, or at least more appropriate for making multiple comparisons.

References

APA; American Psychological Association. 2010. Publication Manual of the American Psychological Association (6th ed.).

Austin, P.C. and J.E. Hux. 2002. A brief note on overlapping confidence intervals. Journal of Vascular Surgery, 36:194–195. doi:10.1067/mva.2002.125015.

Belia, S., F. Fidler, J. Williams, and G. Cumming. 2005. Researchers misunderstand confidence intervals and standard error bars. Psychological Methods, 10: 389–396. doi:10.1037/1082-989X.10.4.389.

Benamini,Y. and H. Braun. 2002. John W. Tukey-Kramer's contributions to multiple comparisons. The Annals of Statistics, 30: 1576–1594. projecteuclid.org/euclid.aos/1043351247.

Cumming, G., F. Fidler, and D.L. Vaux. 2007. Error bars in experimental biology. The Journal of Cell Biology, 177: 7–11. doi:10.1083/jcb.200611141.

Cumming, G., J. Williams, and F. Fidler. 2004. Replication and researchers' understanding of confidence intervals and standard error bars. Understanding Statistics, 3: 299–311. doi:10.1207/s15328031us0304_5.

Franz, V. and G. Loftus. 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.

Knoll, M.J., W.R. Pestman, and D.E. Grobbee. 2011. The (mis)use of overlap of confidence intervals to assess effect modification. European Journal of Epidemiology, 26: 253–254. doi:10.1007/s10654-011-9563-8.

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. doi:10.3758/BF03204506.

Loftus, G.R. and M.E.J. Masson. 1994. Using confidence intervals in within-subjects designs. Psychonomic Bulletin & Review, 1: 476–490. doi:10.3758/BF03210951.

O'Brien, R. and M. Kaiser. 1985. MANOVA method for analyzing repeated measures designs: an extensive primer. Psychological Bulletin, 97(2): 316–333.

Schunn, C.D. 1999. Statistical significance bars (SSB): A way to make graphs more interpretable. http://www.lrdc.pitt.edu/schunn/ssb/SSB.rtf.

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

Tukey, J. 1993. Graphic comparisons of several linked aspects: alternatives and suggested principles. Journal of Computational and Graphical Statistics, 2(1): 1–33.

Wainer, H. 1996. Depicting error. The American Statistician, 50(2): 101–111.

Zar. J. 2010. Biostatistical Analysis, 5th ed. Prentice Hall.

Appendix A

Eight Samples with Similar Variances but Differing Sample Sizes

	Sample							
	A	B	C	D	E	F	G	H
	0	0	1	1	3	3	3	3
	1	1	2	2	4	4	4	4
	2	2	3	3	5	5	5	5
	3	3	4	4	6	6	6	6
	4	4	5	5	7	7	7	7
		0		1	4	3		3
		1		2	5	7		4
		2		3	6			5
		3		4				6
		4		5				7
mean	2	2	3	3	5	5	5	5
variance	2.500	2.222	2.500	2.222	1.714	3.000	2.500	2.222
SEM[a]	0.679	0.480	0.679	0.480	0.537	0.574	0.679	0.480
n[b]	5	10	5	10	8	7	5	10
df[c]	4	9	4	9	7	6	4	9

[a]SEM = standard error of the mean

[b]*n* = sample size

[c]*df* = degrees of freedom.

Appendix B

ANOVA Table for the Samples in Appendix A

Source	Sum of the squares	df	Mean square error	*F*	*P*
Corrected model	101.250[a]	7	14.464	6.268	< 0.001
Intercept	770.642	1	770.642	333.945	< 0.001
Between groups	101.250	7	14.464	6.268	< 0.001
Within groups	120.000	52	2.308		
Total	1065.000	60			
Corrected total	221.250	59			

Note. The output was generated by SPSS except that the *P* values were reported as .000.

[a]r^2 = .458, adjusted r^2 = .385.

Appendix C

Pairwise Comparisons of All Samples in Appendix A. *P* values were calculated with Tukey-Kramer tests.[a]

Pair	*P* value	Pair	*P* value	Pair	*P* value
A vs B	0.999	B vs E	0.003	D vs E	0.124
A vs C	0.966	B vs F	0.005	D vs F	0.155
A vs D	0.928	B vs G	0.015	D vs G	0.261
A vs E	0.022	B vs H	0.001	D vs H	0.084
A vs F	0.028	C vs D	0.999	E vs F	0.999
A vs G	0.055	C vs E	0.308	E vs G	0.999
A vs H	0.015	C vs F	0.341	E vs H	0.999
B vs C	0.928	C vs G	0.440	F vs G	0.999
B vs D	0.818	C vs H	0.261	F vs H	0.999
				G vs H	0.999

[a] Results were obtained from SPSS. Values of 1.000 were changed to 0.999.