Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 21, Number 5, March 2016 ISSN 1531-7714

Confidence Intervals for Effect Sizes: Applying Bootstrap Resampling

Erin S. Banjanovic, *University of Louisville*
Jason W. Osborne, *Clemson University*

Confidence intervals for effect sizes (CIES) provide readers with an estimate of the strength of a reported statistic as well as the relative precision of the point estimate. These statistics offer more information and context than null hypothesis statistic testing. Although confidence intervals have been recommended by scholars for many years, these statistics are often not reported. This may be partially due to the complexity of calculating confidence intervals for many statistics. Bootstrap resampling can be used to easily estimate confidence intervals around almost any type of point estimate. The aim of this paper is to demonstrate this methodology using real-world data and to develop several simple principles around this methodology to guide readers in appropriate application.

Effect sizes (ES), confidence intervals (CI), and confidence intervals for effect sizes (CIES) are indicators of *practical significance*. They are frequently treated as an inferior and less important set of statistics than those used to conduct null hypothesis statistical testing (NHST), which use p-values to inform *statistical significance*. However, NHST only provides a true/false response as to whether the traditional confidence interval includes a null value (exclusion results in statistical significance). CIES can provide the same information. The ES summarizes the magnitude of an effect or the strength of a finding in a standardized manner (e.g., correlation, Cohen's D, odds ratio) and the CI provides information about the precision of a point estimate and the potential generalizability or replicability of the estimate. Taken together, they offer all of the information necessary to conduct NHST and the provide insight into how precisely a researcher has estimated the importance (magnitude) of an effect.

The utility of ES, CI, and CIES was first recognized on an international level by the 1999 APA Task Force on Statistical Inference report (Wilkinson, 1999). The Task Force developed a list of best practices to modernize statistical practice for the 21st century. These best practices discouraged over-reliance on NHST and recommended that researchers report ES and CI as context for statistically significant effects. Shortly after, the fifth edition of APA's *Publication Manual* (2001) was released and it echoed the Task Force's recommendation. In the time that followed, discussion around the use and implementation of ES and CI increased. One notable scholar suggested the Task Force's guidelines be extended to include the reporting of CIES as such statistics promote meta-analytic thinking (Thompson, 2002). Now more than fifteen years after the Task Force's initial recommendation, reporting of ES, CI, and CIES is still not common practice in the behavioral and social sciences.

The infrequent use of CI and CIES may in part be due to estimation difficulty. Traditionally, intricate, multi-step formulae have been used to compute confidence intervals by hand (Cumming & Finch, 2001;

Fidler & Thompson, 2001). These formulae are only available for select effect size statistics and they generally require advanced statistical knowledge to implement. Additionally, many of these formulae require assumptions that might not always be viable. However, as computers and statistical software packages developed, a more computationally intense method of estimating CI has become readily available. This method is known as bootstrap resampling and it can be used to identify a CI for any statistic, regardless of the data's underlying distribution (DiCiccio & Efron, 1996; Efron & Tibshirani, 1994). This method allows CI to be estimated for statistics that are not normally presented with confidence intervals, such as medians, Cronbach's alpha, results from exploratory factor analysis, and common effect sizes (e.g., eta-squared).

The aim of this paper is to describe the utility of bootstrap analysis in calculation of confidence intervals and demonstrate the method with real-world data. Key principles of bootstrapping are highlighted throughout the examples presented. R syntax for each of these analyses is provided in the Appendix. For examples of SPSS syntax for similar analyses (without paying for the bootstrap module), see Osborne (2015, pp. 352-353) and Osborne (2014, pp. 75-76).

Bootstrapping CI: Process and Methods

Bootstrapping Process

Bootstrap resampling is a systematic method of computing CI for nearly any estimate. In most research, you start off with the population of interest, take a sample from the population and run your analyses on that sample. If bootstrapped CI are desired, an additional sub-sampling and replication step is added to the analysis phase. This process starts by taking thousands of 'bootstrapped samples' or 'bootstrapped resamples' from the original sample using random sampling with replacement. This results in thousands of resamples containing the same number of subjects as the original sample, but may contain specific records more than once. The analysis of interest is then replicated in each of these resamples, which leaves you with thousands of estimates of the static of interest. Together, those estimates are known as the bootstrap distribution. This process is summarized in Figure 1. DiCiccio and Efron (1996) recommend that at least 2000 replications are used when conducting bootstrap resampling; however we use 5000 replications throughout this paper as more bootstrapped samples improves estimation and has little downside in terms of processing time (i.e., it takes a modern computer only slightly longer).

The idea behind this method is that the resamples can be viewed as thousands of potential samples from the population. Together, the estimates from the resamples represent the possible range of the estimate in the population. A robust empirical CI can then be estimated from the bootstrap distribution.

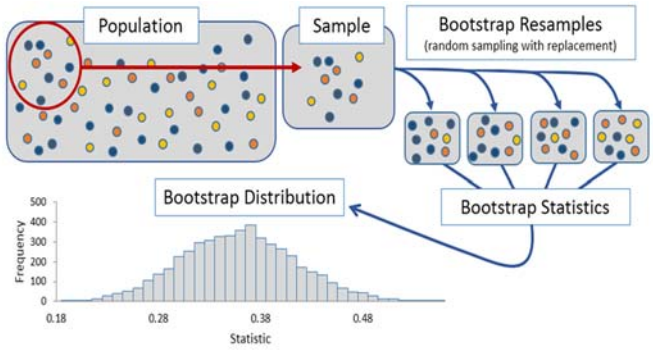

Figure 1. Summary of Bootstrapping Process

Bootstrapping Method

There are many different methods for estimating CI from a bootstrapped distribution. Table 1 summarizes five of the more common methods. These methods use the bootstrap distribution in different ways to arrive at CI. The normal interval method only uses the bootstrap distribution to get an estimate of the standard error (SE), which it then uses in the more traditional CI formula (see Table 1). The remaining methods actually derive the estimate entirely from the bootstrapped distribution. The percentile interval, studentized interval, and BCa interval all conclude with the same step of identifying the percentiles corresponding to the desired CI as the upper and lower bounds (e.g., 2.5% and 97.5% for a 95% CI). These estimates differ however in adjustments made to the bootstrap distribution before this step. The percentile interval makes no adjustments, the studentized interval converts the distribution to studentized statistics, correcting each statistic by its associated standard error, and the BCa interval corrects the distribution for bias and acceleration. Finally, the basic interval method corrects the distribution for bias and then identifies the lower and upper bounds that capture the desired CI level using a slightly more complex formula.

Table 1. Methods Used for Bootstrapped 95% CI Estimation

Method	Description	Assumptions
Normal Interval	The standard error (SE) is computed as the standard deviation (SD) of the bootstrap distribution. The CI are then computed by: $\theta^* \pm 1.96 * SE$, where θ^* is the sample estimate.	– The distribution of the bootstrapped statistic is approximately normal and symmetric. – The sample estimate is an unbiased estimator of the population estimate.
Percentile Interval	The CI are the estimates at the .025 and .975 quantiles of the bootstrap distribution.	– The distribution of the bootstrapped statistic is approximately symmetric. – The sample estimate is an unbiased estimator of the population estimate.
Basic Interval	The CI are estimated by correcting the bootstrap distribution for bias, or skew, and solving for the estimates which capture 95% of the bootstrap statistics.	– The sample estimate is an unbiased estimator of the population estimate.
Studentized Interval	The statistic and SE of the statistic are computed for each of the bootstrap resamples. The bootstrap distribution is transformed into a distribution of studentized statistics[1] and the CI are found at the .025 and .975 quantiles.	– The standard error for the estimate can be computed.
Bias-Corrected & Accelerated Interval (BCa)[A]	The bootstrap distribution is corrected for bias (i.e. skew) and acceleration (i.e., nonconstant variance) and the CI are found at the .025 and .975 quantiles of the corrected distribution.	– None.

[A] This method tends to suffer from convergence problems.
Note. Please see Hall (1992), Efron & Tibshirani (1994), or Davison & Hinkley (1997) for additional information about these methods and Carpenter & Bithell (2000) for a brief review of the methods and assumptions.

The selection of an appropriate bootstrapping method generally can be determined based on responses to fo[1]ur basic questions. These questions target the underlying assumptions of the methods (see Table 1) in order to select one that best fits the data. In general, these assumptions are tested by examining whether we have sufficient information for the method (i.e., standard error) and examining the bootstrap distribution. Remember the bootstrap distribution theoretically reflects the range of possible estimates in the population, thus we can use this information to review for potential problems that can introduce bias into our estimate. Some of the methods are more robust to such issues than others. The four basic questions to review are as follows:

1) Is there a formula to estimate the standard error of the statistic,
2) Is the distribution symmetrical around the mean of the bootstrap resampled statistics,
3) Is the distribution normal/ Gaussian, and
4) Is the sample estimate a biased estimate of the population statistic.

The first three questions are relatively easy to answer. The first asks if we can derive an estimate of the standard error for the statistic of interest. There are a number of statics for which there are no readily available estimates of standard error, and even when they do exist, bootstrap resampling may provide advantages. The second and third questions ask if the bootstrap distribution is approximately symmetrical and /or corresponds reasonably well to a normal/ Gaussian distribution. Finally, the last question targets the

[1] The studentized statistic is computed by $(\theta^* - \hat{\hat{\theta}})/SE$, where $\theta^*$is a bootstrap estimate, $\hat{\theta}$ is the sample estimate, and *SE* is the standard error for the bootstrap sample.

existence of estimator bias. We can test this by comparing the estimate in the original sample to the average estimate in the bootstrapped distribution. While this is not a true comparison to the population estimate, it can provide an indicator of bias that might exist. A demonstration of how to check these assumptions and the differences when assumptions are violated is provided in Example 2 below.

In practice, we have found little difference between these methods. We often default to using the percentile method as it is one of the easiest to implement and more readily grasped by readers. However, we do recommend checking the assumptions of the model before using it, particularly if you have a smaller sample or your variables are known to suffer from non-normality. If you have concerns about which method to use, try running a few of the different methods and see if they produce different results. If they do, use a method that is more robust (e.g., BCa).

Example 1: Mean

Many undergraduate students and graduate students are introduced to CI in an introductory statistics course. They learn how to calculate a 95% CI for a mean by hand, using the formula: $\bar{x} \pm 1.96 * (s/\sqrt{n})$, where $\bar{x}$ is the mean, s is the standard deviation, and n is the sample size[2]. This method of estimating CI for a mean is generally straight-forward and does not require the more complex bootstrapping procedures. However, we will start here as a pedagogical example of applying bootstrap methods to calculation of 95%CI.

Data from the 2010 cohort of the Health and Retirement Study (RAND Center for the Study of Aging, National Institute on Aging, & the Social Security Administration, 2013) were used for this example. A subset containing 13,456 individuals between the ages of 40 and 102, with complete retirement information (the year they retired or planned to retire) were used. The 13,456 individuals in our dataset will be viewed as the 'population' that we are aiming to represent.

Traditional CI vs Bootstrapped CI

First, we compare the CI produced by empirical bootstrap resampling techniques (via the percentile interval with 5000 resamples) to those produced by the traditional 'by-hand' method reviewed above to demonstrate the equivalence. We will take five random samples of 80 individuals from this 'population' to represent different samples that a researcher might collect. We then compute and compare the results (see Table 2). In this example, the two methods are found to produce similar results, with differences generally at the first or second decimal place. It is important to note that these methods will produce similar results unless the data violate parametric assumptions (e.g., data is skewed). In such cases, the bootstrapped CI will yield a better estimate because that calculation does not rely on distributional assumptions of the data (DiCiccio & Efron, 1996).

Table 2. Confidence Interval Estimates for Mean Retirement Age

Sample	Mean	Traditional 95% CI		Empirical 95% CI[A]	
		Lower Bound	Upper Bound	Lower Bound	Upper Bound
1	60.83	58.51	63.14	58.47	63.05
2	61.14	59.36	62.91	59.33	62.87
3	63.40	61.46	65.34	61.39	65.30
4	61.35	58.91	63.79	58.91	63.74
5	62.60	60.83	64.37	60.82	64.31

[A] The empirical 95% CI were computed using the percentile method. This will be discussed more thoroughly in example 2.

Variability Across Samples

Next, we took additional random samples from the population to demonstrate the variability in means and empirical CI among samples from the same population. We will take an additional 15 samples, each containing 80 individuals, to provide a total of 20 samples from the population. Note, we are purposefully selecting a smaller sample size (that researchers still routinely use) as they are known to be more volatile in estimating population parameters. The empirical CI (estimated by the percentile interval with 5000 resamples) for each of the 20 samples are presented in Figure 2. The 'population' mean is presented as a solid blue line.

[2] The 1.96 constant in the formula is the z-score corresponding with the 95% CI.

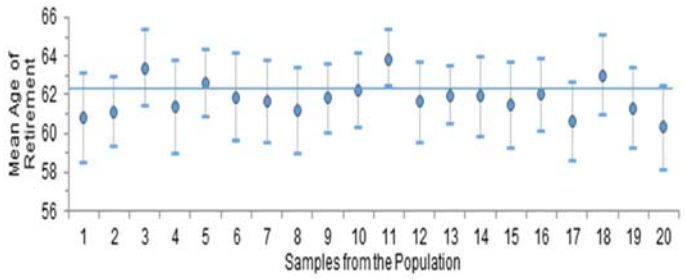

Figure 2. Mean Retirement Age (Actual and Expected) and Empirical 95% CI

As you can see in Figure 2, the "population" mean retirement age is approximately 62 years of age. Although there was a good amount of volatility in the point estimate (the mean), nineteen of the twenty samples contained the population mean within their bootstrapped confidence interval estimates, further verifying that we can generally be 95% certain that our CI contain the population mean. Most CIs are reasonably narrow, also indicating a reasonable level of precision and likelihood that replications would result in a reasonably similar outcome.

Principle 1: Bootstrapping cannot fix small or biased samples.

The results in Figure 2 also demonstrate an issue common to small and biased samples: the sample mean may deviate from the population mean. The mean retirement age in sample 11 is 63.9 years of age, approximately two years past the "population" mean, and the 95% CI around the mean does not contain the mean retirement age in the population. If this sample was the only one collected it would misrepresent the population, although the population mean is very close to the lower bound of the CI. However, data cleaning prior to bootstrap analysis can help if influential cases are the cause of the mis-estimation (Osborne, 2015).

Example 2: Median

The median is a better indicator of central tendency than the mean when data are not normally distributed. No simple, distribution-free, formula exists to calculate CI for the median, yet these are easily calculated by using bootstrap methods. This example demonstrates the process of bootstrapping CI around an estimate of the median and introduces the different methods used for bootstrapping CI.

The current example uses a subsample (n=200) of data from the National Health and Nutrition Examination Survey (NHANES; CDC & NCHS, 2010). We are interested in estimating the median and the 95% CI for four variables which are known to have skewed distributions: systolic blood pressure, diastolic blood pressure, body mass index, and average time making dinner. As we mentioned in the first example, bootstrapped CI estimation can yield better estimates than more traditional methods because it does not require the data to meet assumptions of normality.

Bootstrap Methods

To determine which methods are appropriate for our data, we must answer the four basic questions pertaining to the assumptions that were reviewed above: 1) Is there a formula to estimate the SE? 2) Is the bootstrap distribution symmetric? 3) Is the bootstrap distribution normal/ Gaussian? and 4) Is the sample estimate a biased estimate of the population statistic? First, we know there is no true formula for the SE of the median. One method, the studentized interval, requires a formula for SE (see Table 1), thus it can automatically be removed from consideration. Next, we must evaluate the histograms of the bootstrapped statistics, or the bootstrap distributions, to address the remaining questions and determine which methods might be most appropriate for the data. These histograms are presented in Figure 3. Each contains a solid orange line representing the median of the original sample and a solid blue line representing the averaged bootstrapped median.

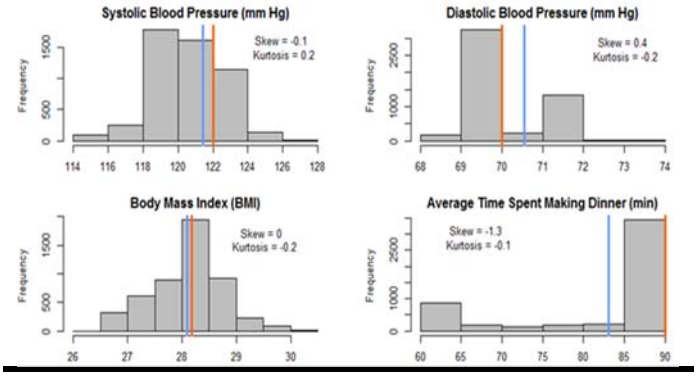

Figure 3. Histograms of Bootstrapped Statistics

Note. The median of the original sample is presented as a solid orange line and the averaged bootstrapped median is presented as a solid blue line..

In reviewing the bootstrap distributions, both systolic blood pressure and body mass index appear to have normal, symmetrical distributions. The sample median and average bootstrapped median are very close for both variables, indicating the sample medians are unbiased estimators of the population medians. Based on these results, we can conclude that the normal, percentile, basic, or BCa interval methods are most appropriate for these two variables (see Table 1). Diastolic blood pressure and average time making dinner both have non-normal, asymmetric distributions. The sample median for diastolic blood pressure is close to the average bootstrapped median, indicating the sample estimate is unbiased. However, the sample median for time making dinner is not very close to the average bootstrapped median, suggesting the sample estimate is a biased estimator. These results suggest the basic or BCa are the most appropriate methods for Diastolic blood pressure and the BCa is the most appropriate for average time making dinner.

Different methods of bootstrap CI estimation are more appropriate than others for particular variables. If we were conducting this analysis for research and found some of the variables to violate particular assumptions, we would likely use the method most appropriate for the majority of variables or statistics of interest (in this case, the BCa method). However, for our pedagogical purpose, we will plot them all and compare. Figure 4 presents the CI for each of the variables (using 5000 bootstrap resamples). The sample median is presented as a solid horizontal line and the CI deemed appropriate are presented in blue.

Depending on the method used, the confidence in the effect estimate may differ particularly when working with non-normal distributions that violate the necessary assumptions. For example, systolic blood pressure and BMI are normally distributed and provide consistent CIs across methods, whereas diastolic blood pressure and time for dinner did not. In the case of diastolic blood pressure, the normal interval method tends to overestimate both the lower and upper bound of the CI, while the percentile method tends to underestimate the lower bound and overestimate the upper bound. Finally, in regard to time for dinner, the normal and basic methods greatly overestimate the upper bound of the CI.

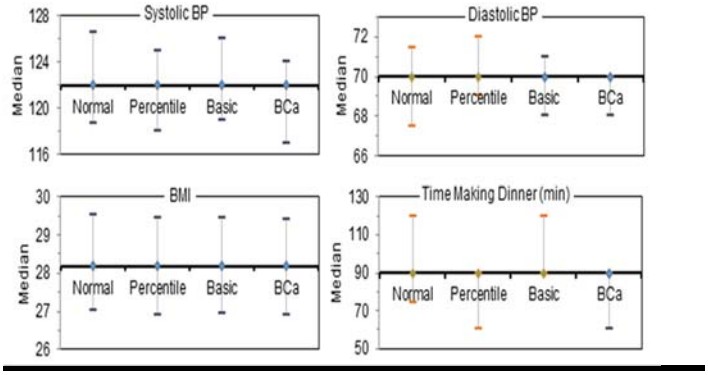

Figure 4. Empirical CI for Median Blood Pressure, BMI, & Time to Make Dinner

Note. CI whose assumptions were met by a variable are presented in blue. CI whose assumptions were not met are presented in orange. Upper boundaries in the right two boxes were skewed, leading to the mean and upper bound being so close as to not be distinguishable.

Principle 2: Different methods of bootstrapping CI can yield different results.

These results highlight that the method used to estimate CI can make a difference. Before selecting a method, the assumptions of the methods should be reviewed and tested. If multiple methods can be used, then it may be worthwhile to review and compare the estimates. If differences exist, it may be advisable to use the method that is most closely aligns with the necessary assumptions.Step 5: Comparing Balance

Example 3: Correlation

Correlations are standardized measures of association. There are formulae to calculate CI for correlation coefficients, but they are rarely reported. This example illustrates the estimation of CI for correlations and highlights the effect of sample size on the precision of the estimated CIs.

Data from the Education Longitudinal Study of 2002 (U.S. Dept. of Education & National Center for Education Statistics, 2010) are used to examine the relationship between student math achievement and three other variables: reading achievement, socio-economic status (SES), and student belief that "Math is

Fun". The 'population' for this example includes 10,353 10th grade students that were surveyed in 2002.[3]

Sample Size

Five random subsamples of the data, ranging in size from 50 to 1500, are taken to represent five samples of varying sizes that a researcher could take from the population. Correlation coefficients and empirical CI (via the percentile method with 5000 resamples) are then estimated for each sample in order to demonstrate how sample size can impact the precision of an estimate and the use of CI in interpreting this precision.

Figure 5 summarizes the results of this analysis. The correlation in the 'population' of 10,353 students is presented as a solid blue line. The correlation in each of the samples from the population is presented as a blue dot. If we examine the figure, we find math achievement is most highly correlated with reading achievement, followed by SES and the belief that "Math is Fun". We also see the samples with smaller sample sizes are generally associated with larger CI and estimates that are further from the estimate in the population4. This leads us to our third principle.

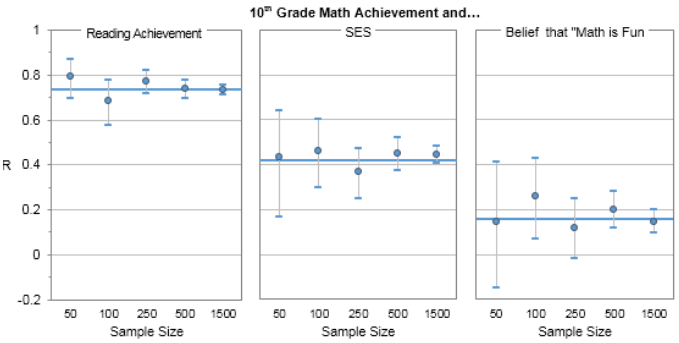

Figure 5. Correlations and their Empirical 95% CI for 5 Different Sample Sizes

Note. The empirical 95% CI were computed using the percentile method

Principle 3: Precision of the point estimate and CI improve with increases in sample size.

CI becomes narrower as sample size increases. Thus, larger samples will have the narrower confidence intervals, while smaller samples will have wider intervals. This suggests that we have a better estimation of the true value of the correlation in the population when a larger sample size is used. We are more likely to question how well the effect would replicate when a small sample is used and large CI are exhibited.

Example 4: Cronbach's Alpha

Cronbach's Alpha is a commonly used indicator of reliability. It tells us the degree to which a set of items exhibit internal consistency and it is known to be sample specific. It is recommended practice to report Cronbach's Alpha in any study using a set of items to construct a scale. Estimation of CI around alpha estimates can further bolster the use of alpha by providing an indicator of how these results might generalize.

Data from The National Science Foundation Surveys of Public Understanding of Science and Technology, 1979-2006 (Miller, Kimmel, ORC Macro, & NORC, 2009) are used to illustrate the use of CI in estimation of Cronbach's Alpha. A subset of the data containing 313 individuals between the ages of 18-24 are used. Responses to six items, rated on a Likert-type agreement scale from 1(strongly disagree) to 4 (strongly agree), are evaluated as indicators of stereotypes of scientists. For the purpose of this example, we will pretend we created this scale and would like to examine how it functions.

Item-Total Correlations

Item-total correlations represent the correlation between an item and all the other items, where the total of the other items is achieved by summing or averaging them. This is a good place to start when examining scale function because it allows us to see which items exhibit internal consistency with the construct being measured. The item-total correlations for the six items, along with

[3] Sample weighting was not applied as we are not using the data to draw substantive conclusions., but rather to demonstrate bootstrap analyses.

[4] Note, two of the correlations in the smallest sample are actually quite close to the population estimate. However, the CI suggest that there was great variability in the estimate among the bootstrap resamples and thus it is difficult to have confidence in the estimate.

95% empirical CI (via the percentile method with 5000 resamples), are presented in Figure 6. In general, the item-total correlations are fairly low, ranging from 0.13 to 0.53 and the CI tend to be quite large. These results suggest we are likely to see variability in these estimates if additional samples are collected and the estimates will likely remain lower than desired. Furthermore, one item is found to have a CI that includes zero, suggesting the item may contribute very little or potentially nothing to our construct of interest. Overall, these initial results are discouraging and we might stop here if our goal was simply to evaluate these items potential as a scale; but it is not so let's continue**.**

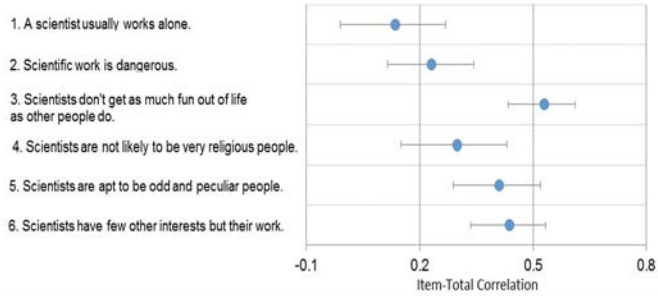

Figure 6. Item-Total Correlations and Empirical 95% CI for 6 Items

Note. The empirical 95% CI were computed using the percentile method.

Reliability

Cronbach's alpha and 95% empirical CI (via the percentile method with 5000 resamples) were estimated for the six items. The alpha was rather low, α=0.60, and the CI ranged from 0.53 to 0.67. Together, these results lead us to expect that there will be significant variability in the point estimate for alpha in other similar samples, and that none of them are expected to rise to a level one could consider adequate.

Principle 4: CI can provide insight on generalizability.

If we were analyzing these results without the CI we could get a different view of this scale. Even though the Cronbach's Alpha estimates are lower than desired, we might convince ourselves that the results represent a lower bound estimate and that we could see better results in a different sample. The upper bound of the CI estimates does not exceed .7, thus they have disabused us of such thoughts. In this example, the CI tell us that the scale has relatively low internal consistency among the population of interest but the differences between the groups are likely to replicate. In this way, the CI provide us a lens through which to glimpse the potential generalizability of the scale to the intended population.

Example 5: Repeated Measures ANOVA

A central hallmark of science is replication (Killeen, 2008; Thompson, 2002; Yu, 2003). While not perfect, resampling methods can inform the researcher as to the potential replicability or non-replicability of an effect. This fifth example explores the use of bootstrapping CI as an indicator of power.

This example uses data from a colleague performing a study in a physical therapy environment in Australia (Gabel, Osborne, & Burkett, 2015). The study aimed to compare a new physical therapy technique for knee rehabilitation (#5) to four standard ones using a measure of muscle activity captured via electromyography (EMG). At the time of initial analysis, the researcher had 21 participants, with plans to measure a total of 35. As this type of data collection is time consuming, he wanted to know if a sample of 35 would have sufficient power. The initial results, summarized Table 3, were promising and highly significant ($F(4,80) = 11.60$, $p < .0001$, $\eta 2 = 0.65$).[5]

Table 3. EMG measures in knee rehabilitation patients across five activities (N=21)

			Traditional 95% CI	
EMG	Mean	SE	Lower Bound	Upper Bound
1	82.71	14.94	54.44	111.99
2	91.14	15.92	59.95	122.33
3	97.10	13.46	70.71	123.48
4	90.57	13.93	63.26	117.88
5	141.38	17.53	107.02	175.74

Despite the small sample, the first four conditions are relatively similar, and the fifth is markedly different in terms of effort as measured by EMG. It is possible to calculate power in simple repeated measures ANOVA

[5] Greenhouse-Geisser corrected statistics reported due to lack of sphericity.

such as this, but the real question from the colleague was whether he could trust that this effect was real. In other words, as a physical therapist he is probably more interested in whether his results will generalize to other patients or whether this effect is unstable. Although external replication is ideal (replication with an independent sample), internal replication analyses can provide insight by estimating the probabilities of replication given other similar samples. Additionally, bootstrap resampling can be used to estimate the expected stability of effect size estimates (in this case, eta-squared).

Internal Replication

Bootstrap analysis with 5000 resamples attempted to shed light on the probability of another sample replicating these results. As Table 4 shows, the average expected *F* is 13.57 (CI: 0.00, 17.27), which is a wide interval, not unexpected with such a small sample. However, even with such a small sample, 100% of the bootstrapped resamples yielded a significant effect, which suggested that the effects were likely to replicate. Further, the CI around the eta-squared estimate were large, ranging from 0.57 to 0.79. Thus, the expectation was that a similar sample would produce similar results (i.e., a reasonably strong effect) and thus was worth pursuing. Indeed, the researcher subsequently gathered another data set that showed this bootstrapped prediction was a very good estimate of the effect in an independent (replication) sample.

Principle 5: CI can predict replication

Estimation of CI for tests of significance and effect sizes can predict the replication of the significance and magnitude of an effect. If CI for tests of statistical significance only contain values with associated p-values below .05, the results are likely to replicate within a similar sample. If the CIs contain values associated with p-values greater than .05, the results are not as likely to replicate. Similarly, if CIES contain a relatively small range that primarily contains small, medium, or large effect sizes, we are likely to see similar results in other samples. If the CIES contain a larger range, replication is less predictable.

Example 6: Logistic Regression

Logistic regression is performed to understand the variables that contribute to a binary outcome variable, such as whether students in high school will graduate or not. The current example uses to data from the National Education Longitudinal Study of 1988 (U.S. Dept. of Education & National Center for Education Statistics, 1989) to predict graduation based on student socio-economic status (SES). Five random samples of 200 students were taken from the dataset to demonstrate trends in CI estimation that can inform the utility of an estimate.

Logits, Odds-Ratios, and Predicted Probabilities

Table 5 displays the results for logistic regressions to predict graduation in each sub-sample. The likelihood ratio test and the SES coefficient in each of the models are significant. These results tell us that each model is significantly improved by inclusion of SES. We can then proceed to examine the effect of SES on graduating. While the results of logistic regression are generally output as logits, we convert them to odds ratios as this is a more intuitive metric to understand. The odds ratios for SES range from 1.72-3.97 across the models. This can be interpreted to mean the odds of graduating increase by 1.72-3.97 as the SES levels rise by 1 standard deviation.

Table 4. Summary of bootstrap analysis of EMG data.

		Bootstrapped Distribution					95% Empirical CI	
Statistic	N	Mean	Median	SD	Minimum	Maximum	Lower Bound	Upper Bound
F	5000	13.49	12.84	4.83	2.35	43.47	0.00	17.27
P-value	5000	0.00	0.00	0.00	0.00	0.06	0.00	0.00
Significant flag	5000	1.00	1.00	0.02	0.00	1.00	1.00	1.00
Partial Eta^2	5000	0.56	0.56	0.08	0.29	0.79	0.57	0.79

Note. The empirical 95% CI for the F stat, p-value, and significant flag were computed using the basic method while the CI for the eta-squared was computed using the BCa method. Methods were identified based on assumptions and different methods were used because the BCa method suffered from extreme endpoints for some of the variables.

Table 5. Odds-Ratios for Logistic Regression to Predict Graduating in the Next Hour

	Logit		Odds Ratio		
Sample	Intercept	SES [A]	Intercept	SES [A]	Likelihood Ratio Test
1	2.56***	1.16**	12.89	3.19	14.38***
2	2.47***	0.54*	11.85	1.72	4.45*
3	2.50***	1.00***	12.17	2.71	3.84***
4	2.66***	1.38***	14.28	3.97	29.35***

[A] SES was standardized to have a mean of zero and a standard deviation of 1 among those in the NELS: 88 dataset.
* p<.05. ** p<.01. *** p<.001.

While odds ratios are more intuitive than logits, there is a metric that is even easier to understand: predicted probability. Predicted probabilities represent the probability of an event occurring based on a set level of a predictor variable. We converted the results in Table 5 to predicted probabilities of graduating for three different SES levels. These predicted probabilities and their empirical 95% CIs (via the percentile method) are presented in Figure 7. Note that we could have estimated CIs for any of the estimates reported in Table 5.

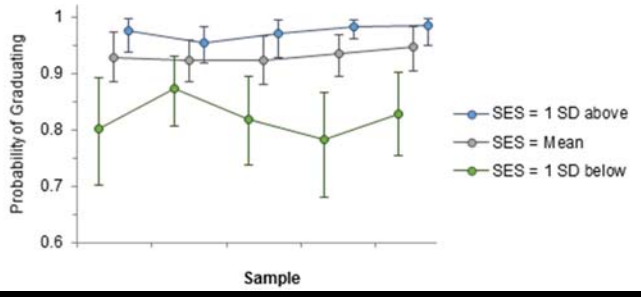

Figure 7. Predicted Probability of Graduating at Three Different SES Levels

Note. The empirical 95% CI were computed using the percentile method.

Figure 7 clearly depicts the relationship between SES and graduating. The probability of graduating ranges from 78-87% at SES levels 1 standard deviation below the mean, 92-95% at mean SES levels, and 95-98% at SES levels 1 standard deviation above the mean. Notice the CI range decreases as SES increases. This can be interpreted to mean that when SES is higher we are generally good at predicting that graduation will occur. However, as SES level decreases and graduation becomes somewhat less probable, we are less able to accurately predict it. In other words, there is more error associated with the graduation estimates among those at lower SES levels.

Psuedo-R2

Before concluding this example, let us consider one more statistic sometimes reported in logistic regression: the pseudo R^2. Pseudo R^2 is an estimate of the amount of variance explained by a model and an indicator of goodness-of-fit.[6] There is considerable debate over which of the dozen or so methods to use (Menard, 2000; Mittlbock & Schemper, 1996), or even whether to use any of them at all (Osborne, 2015). We proceed by estimating three of the most popular estimates attributed to McFadden (1974), Cox and Snell (1989), and Nagelkerke (1991) to demonstrate often volatile nature of such estimates, and how CIs can make the reader more informed of the precision of an estimate such as these. These estimates and their corresponding CI are presented in Figure 8.

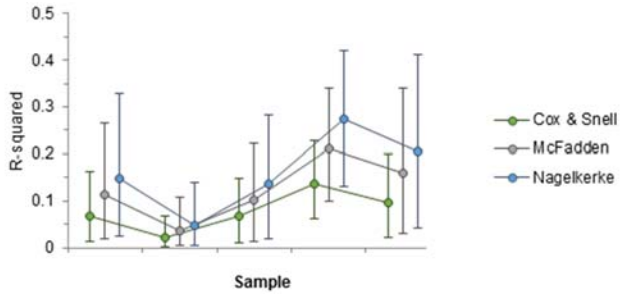

Figure 8. Three Pseudo R2 Estimates and their Empirical CI

Note. The empirical 95% CI were computed using the BCa method.

The CI provide insight into the error assocated with each estimate. If we examine the psuedo R2 estimates alone, we find the Cox and Snell estimates provide the smallest estimates while the Nagelkerke provide the highest. When we examine the CI we find the Nagelkerke estimates, that provide the most favorable indication of association, are also paired with the largest error. The Cox and Snell estimates provide the smallest

[6] It is, of course, a *pseudo* R^2 for a reason- in logistic regression, which utilizes maximum likelihood estimation, we do not directly measure variance accounted for.

estimate of association, but they are linked with the least error. While the Cox and Snell estimates may provide the most consistent estimates out of these methods, the results do not indicate a strong likelihood of replication of the point estimate.

Principle 6: CI can inform us of the utility of an estimate.

Large CI indicate there is a greater degree of error in an estimate and that the results may not replicate or generalize. In the current example, the large CI around the predicted probabilities of graduating for low SES students suggest the point estimate is not very reliable and should be interpreted with caution. Similarly, when large CI are consistently associated with a particular type of estimate or effect size this can inform our understanding of the reliability of the estimate itself.

Conclusions

Good science must be informed by replication, although in the social and behavioral sciences, replication is not currently common. Precision is also important, yet often overlooked in publications. Confidence intervals can help inform readers as to how precise point estimates are, and how likely the point estimates are to be replicated in another similar sample. Although authors are encouraged to publish effect sizes and confidence intervals, there is a methodoligical gap in estimation of these CIs for certain statistics. Our goal in this paper was to demonstrate an easily accesible, empirical method for estimating confidence intervals for these (or almost any other) statistics using bootstrap analysis (e.g., Osborne, 2014 demonstrates the use of bootstrap analysis for evaluating the stability of factor loadings and eigenvalues in exploratory factor analysis).

In this paper we present examples of applications of bootstrap analysis for producing CIs for statistics such as medians, Cronbach's alpha, partial eta squared, and pseudo-R2 statistics. We also demonstrated some basic principles that researchers should keep in mind when conducting bootstrap analysis. For example:

1) bootstrap analysis cannot repair a fatally flawed (e.g., highly biased) sample,
2) different methods of calculation of bootstrap-based CIs require different assumptions,
3) larger samples tend to produce better precision and narrower confidence intervals
4) confidence intervals can inform inferences about the possible generalizability of results
5) confidence intervals can predict replication, and
6) bootstrapped confidence intervals can reveal important information about the utility of a statistic.

The methods presented in this paper provide researchers with the tools to routinely report confidence intervals for effect estimates for a broad range of statistical estimates. If embodied, these methods would provide readers with context and estimates of precision for commonly reported statistics.

References

American Psychological Association. (2001). Publication manual of the American Psychological Association, 5th Edition. Washington, D.C.: American Psychological Association.

Carpenter, J. and Bithell, J. (2000), Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19: 1141–1164.

Centers for Disease Control and Prevention (CDC), & National Center for Health Statistics (NCHS). (2010). National Health and Nutrition Examination Survey Data. Retrieved from: http://www.cdc.gov/nchs/nhanes.htm

Cox, D. R., & Snell, E. J. (1989). Analysis of Binary Data (2nd ed.): Chapman & Hall.

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(4), 532-574.

Davison, A. C., & Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge, United Kingdom: Cambridge University Press.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap Confidence Intervals. Statistical science : a review journal of the *Institute of Mathematical Statistics.*, 11(3), 189-212.

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap: Chapman & Hall/CRC.

Fidler, F., & Thompson, B. (2001). Computing Correct Confidence Intervals for ANOVA Fixed- and Random-Effects Effect Sizes. *Educational and Psychological Measurement*, 61(4), 575-604.

Gabel, C. P., Osborne, J. W., & Burkett, B. (2015). The influence of slacklining on quadriceps rehabilitation,

activation, and intensity. *The Journal of Science and Medicine in Sport,* 18, 62-66.

Hall, P. (1992). The bootstrap and Edgeworth expansion (Springer series in statistics). New York: Springer-Verlag.

Killeen, P. R. (2008). Replication Statistics. In J. W. Osborne (Ed.), Best Practices in Quantitative Methods (pp. 103-124). Thousand Oaks CA: Sage.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in Econometrics (pp. 105-142).

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54, 17-24.

Miller, J. D., Kimmel, L., ORC Macro, & NORC. (2009). National Science Foundation Surveys of Public Attitudes Toward And Understanding of Science And Technology, 1979-2006 [Computer file]. 3rd Roper Center version. Storrs, CT: Roper Center for Public Opinion Research. Retrieved from http://www.ropercenter.uconn.edu/data_access/data/datasets/nsf.html#download_data

Mittlbock, M., & Schemper, M. (1996). Explained variation in logistic regression. Statistics in Medicine, 15, 1987-1997.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691-692.

Osborne, J. W. (2014). Best practices in Exploratory Factor Analysis: Createspace Publishing.

Osborne, J. W. (2015). Best practices in logistic regression. Los Angeles: SAGE.

RAND Center for the Study of Aging, National Institute on Aging, & the Social Security Administration. (2013). RAND HRS Data file. Retrieved from: http://www.rand.org/labor/aging/dataprod/hrs-data.html

Thompson, B. (2002). What Future Quantitative Social Science Research Could Look Like: Confidence Intervals for Effect Sizes. *Educational Researcher*, 31(3), 25-32.

U.S. Dept. of Education, & National Center for Education Statistics. (1989). National Education Longitudinal Study, 1988 [Computer file]. Washington, DC. U.S. Dept of Education, National Center for Education Statistics. Retrieved from http://nces.ed.gov/surveys/nels88/

U.S. Dept. of Education, & National Center for Education Statistics. (2010). Education Longitudinal Study (ELS), 2002: Base Year [Computer file]. Washington, DC. U.S. Dept of Education, National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2010338

Wilkinson, L. (1999). Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, 54, 8,594-604.

Yu, C. H. (2003). Resampling methods: concepts, applications, and justification. *Practical Assessment, Research & Evaluation*, 8(19), 1-23. Retreived from http://pareonline.net/getvn.asp?v=8&n=19

Citation:

Banjanovic, Erin S., and Osborne, Jason W. (2016). Confidence Intervals for Effect Sizes: Applying Bootstrap Resampling. *Practical Assessment, Research & Evaluation*, 21(5). Available online: http://pareonline.net/getvn.asp?v=21&n=5.

Corresponding Author

Erin S. Banjanovic
700 North Hurstbourne Parkway, Suite 100
Louisville, KY 40222-5393

erin.smith.3 [at] louisville.edu