Behav Res (2017) 49:294–309
DOI 10.3758/s13428-016-0702-8

Confidence intervals for correlations when data are not normal

Anthony J. Bishara[1] · **James B. Hittner**[1]

Published online: 28 January 2016

Abstract With nonnormal data, the typical confidence interval of the correlation (Fisher z') may be inaccurate. The literature has been unclear as to which of several alternative methods should be used instead, and how extreme a violation of normality is needed to justify an alternative. Through Monte Carlo simulation, 11 confidence interval methods were compared, including Fisher z', two Spearman rank-order methods, the Box–Cox transformation, rank-based inverse normal (RIN) transformation, and various bootstrap methods. Nonnormality often distorted the Fisher z' confidence interval—for example, leading to a 95 % confidence interval that had actual coverage as low as 68 %. Increasing the sample size sometimes worsened this problem. Inaccurate Fisher z' intervals could be predicted by a sample kurtosis of at least 2, an absolute sample skewness of at least 1, or significant violations of normality hypothesis tests. Only the Spearman rank-order and RIN transformation methods were universally robust to nonnormality. Among the bootstrap methods, an observed imposed bootstrap came closest to accurate coverage, though it often resulted in an overly long interval. The results suggest that sample nonnormality can justify avoidance of the Fisher z' interval in favor of a more robust alternative. R code for the relevant methods is provided in supplementary materials.

Electronic supplementary material The online version of this article (doi:10.3758/s13428-016-0702-8) contains supplementary material, which is available to authorized users.

✉ Anthony J. Bishara
BisharaA@cofc.edu

[1] Department of Psychology, College of Charleston, 66 George Street, Charleston, SC 29424, USA

Keywords Correlation · Confidence interval · Normal · Robust · Fisher z · Fisher z' · r to z

Major psychological organizations and journals have recently taken a stand on an issue in statistics: They have endorsed more frequent use of confidence intervals (American Psychological Association, 2010; Lindsay, 2015; Psychological Science, 2014; Psychonomic Society, 2012). Abiding by such endorsements should be easy, because numerous resources have explained the construction of confidence intervals (e.g., Cumming, 2012), especially for situations in which parametric assumptions are satisfied. Unfortunately, parametric assumptions are rarely met in actual behavioral data, and particularly assumptions about normality. Normality appears to be the exception rather than the rule (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013; Micceri, 1989). Violations of normality might be especially problematic for inferences based on correlations, for which even large sample sizes are unlikely to help (see Hawkins, 1989). In other words, one should not expect to be rescued by the central limit theorem, at least not for the confidence interval of a correlation. When nonnormality is present, the typical parametric confidence interval of the correlation may be inaccurate, and it is not clear which of several alternative methods should be used. The goal of the present research is to systematically compare several alternatives and determine whether sample statistics can be used to inform the choice among these alternatives.

The 95 % confidence interval is an interval that, if constructed in a large number of samples, should cover the true population parameter in 95 % of those samples. The definition of a confidence interval may seem drab, but the advantages of it are not. It has long been suspected that confidence intervals can mitigate some of the well-known limitations of null

hypothesis significance testing (see, e.g., Cohen, 1994; Cumming & Finch, 2005; Loftus, 1996; Tryon, 2001), and this suspicion has been supported by recent studies in statistical cognition. Confidence interval formats have been shown to reduce the likelihood of "accepting the null" (Fidler & Loftus, 2009; Hoekstra, Johnson, & Kiers, 2012), a mistake that commonly frustrates statistics instructors. Additionally, confidence interval formats have been shown to improve the interpretation of multiple studies that diverge on statistical significance but converge on effect direction (Coulson, Healey, Fidler, & Cumming, 2010). Generally, confidence intervals are beneficial because they shift readers' focus toward the continuous dimension of effect size rather than the simple binary dimension of "significant" versus "nonsignificant," and they highlight the idea that estimates of effect size are always uncertain (Cumming, 2012; Fidler & Loftus, 2009; Hoekstra et al., 2012). Thus, there are advantages to using confidence intervals, so the accurate construction of such intervals is important.

For the Pearson correlation coefficient, the default method of constructing a confidence interval is the *Fisher z′* method (Fisher, 1915, 1921). This method is sometimes referred to as *r-to-z* or *r-to-z′ transformation*. First, the Pearson correlation coefficient is calculated as usual:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}. \quad (1)$$

Then, the Fisher z' transformation of it is defined as

$$z' = .5*\ln\left(\frac{1+r}{1-r}\right). \quad (2)$$

The 95 % confidence interval for z' is

$$z' \pm 1.96*\sigma_{z'}, \quad (3)$$

where $\sigma_{z'}$ is the approximate standard error of z':

$$\sigma_{z'} = 1\Big/\sqrt{N-3}. \quad (4)$$

Finally, the upper and lower bounds of the confidence interval are converted back to the r scale using the equation

$$r = \frac{\left(e^{2z'}\right)-1}{\left(e^{2z'}\right)+1}. \quad (5)$$

For example, if $r = .50$ and $n = 40$, the 95 % confidence interval is .22 to .70. Note that z' is a transformation of r, not of the raw data. In other words, the Fisher z' method does not alter the original scale of measurement for X or Y.

Nonnormality

A key assumption of the Fisher z' confidence interval method is that X and Y have a bivariate normal distribution in the population. When this assumption is met, the Fisher z' method is quite accurate, and alternative formulations for the confidence interval usually fail to improve upon Fisher's original method (Fouladi & Steiger, 2008; Puth, Neuhäuser, & Ruxton, 2014).

However, when the bivariate normality assumption is not met, there is little guidance from the methodological literature as to how to proceed. It may be tempting to still use the Fisher z' method, and to rely on the central limit theorem as a solution to the nonnormality. Along these lines, one famous textbook has suggested that "perhaps the safest course is to require rather larger samples in uses of this test when the assumption of a bivariate normal population seems questionable" (Hays, 1994, p. 650). This suggestion could be problematic for at least two reasons. First, even in situations in which the central limit theorem applies, and larger sample sizes allow for an approximation of normal sampling distributions, alternative statistics can still be more efficient and more powerful. Second, the central limit theorem does not readily address the problem of nonnormality when it comes to the Fisher z' method. If data are not bivariate normal, even as n approaches infinity, the Fisher z' method can fail to converge on the proper answer, because its asymptotic standard error is different from the standard error used in Eq. 4 (Gayen, 1951; Hawkins, 1989). Thus, a large sample size will not necessarily address normality violations for the Fisher z' method. Indeed, as will be shown later, increasing sample size can sometimes worsen the performance of this method. For this reason, when analyzing correlations, nonnormality should be carefully considered and measured.

Nonnormality is often measured through skewness and kurtosis. *Skewness* refers to an asymmetry (e.g., a positive skew often indicates a long tail to the right of a distribution). *Kurtosis* refers to tail-weight (Westfall, 2014). Distributions with positive kurtosis (e.g., t distributions) have tails heavier than in the normal distribution, and distributions with negative kurtosis have lighter, less influential tails. Skewness and kurtosis values together do not completely define all possible types of nonnormality, because no two parameters can do so. However, skewness and kurtosis are particularly useful characteristics to focus on, because they are commonly examined and reported by researchers. Additionally, skewness and kurtosis are useful because specific significance tests are associated with them for both univariate and multivariate normality (D'Agostino, Belanger, & D'Agostino, 1990; Mardia, 1970).

It is well-known that kurtosis affects the Pearson correlation coefficient, but the impact of skewness is not as clear. Early work showed that, when the correlation was high, even a small change in kurtosis could have a large impact on the variance of r, and this held true even for larger sample sizes

(Haldane, 1949). Since then, several researchers have emphasized kurtosis (Bishara & Hittner, 2015; Bonett & Wright, 2000; Duncan & Layard, 1973), and some have even suggested that kurtosis is more important than skewness (DeCarlo, 1997; Gayen, 1951; Mardia, Kent, & Bibby, 1979, pp. 148–149). Indeed, even when skewness is zero, Monte Carlo studies have shown that high kurtosis can inflate Type I and II error rates (Bishara & Hittner, 2012; Edgell & Noon, 1984; Hayes, 1996; Puth et al., 2014), increase bias (Bishara & Hittner, 2015; Zimmerman, Zumbo, & Williams, 2003), and reduce the coverage rates of confidence intervals (Puth et al., 2014). Thus, kurtosis clearly has an effect, but controlling for that, what is the effect of skewness? Comparisons of kurtosis and skewness are complicated by the fact that kurtosis is bounded below by skewness, and so extremely skewed distributions will necessarily have high kurtosis. To determine whether skewness has effects in addition to those of kurtosis, both skewness and kurtosis would need to be orthogonally manipulated, something that has yet to be done in this literature.

Raw data transformation methods

The earlier described Fisher z' method involved transformation of the sampling distribution of the correlation, but what if the raw data were also transformed? In other words, one approach to nonnormality in raw X and Y variables is to transform them directly and then proceed as usual, calculating the Pearson correlation and using the Fisher z' to construct the confidence interval. After any raw data transformation, the correlation no longer represents a linear relationship on the original scales of measurement. However, raw data transformation methods can still be useful. The correlation still indicates the strength and direction of the monotonic relationship—that is, the more general relationship that X tends to increase as Y increases or to increase as Y decreases. Characterizing this monotonic relationship is often sufficient, because many theories in psychology lack the specificity to predict a linear relationship. Additionally, even when theories do have such specificity, the measurements might not (Blanton & Jaccard, 2006). Thus, raw data transformation methods can often be useful. Four such raw data transformation methods are considered here: two methods for the Spearman rank-order transformation, a Box–Cox transformation, and a more general, rank-based inverse normal (RIN) transformation.

The Spearman rank-order correlation—commonly used for ordinal variables and for some nonlinear relationships—is also recommended as an alternative to the Pearson correlation when normality is violated (e.g., Field, 2000; Pagano & Gauvreau, 2000; Rosner, 1995; Triola, 2010). The Spearman rank-order correlation can be thought of as a Pearson correlation following transformation into a flat distribution of ranks (i.e., the histogram of the ranks will be flat so long as there are no ties). Because the ranks of X and Y are flat, an alternative variance term is needed for the Fisher z' statistic. We consider the two most popular variance terms, one by Fieller, Hartley, and Pearson (1957) and one by Bonett and Wright (2000), referred to here as Spearman$_F$ and Spearman$_{BW}$, respectively. These confidence intervals have been extensively compared to several alternatives for ordinal data (Ruscio, 2008; Woods, 2007). It is less clear, though, how these methods fare relative to alternatives when using the Spearman correlation as a solution to nonnormality in continuous data. At least in bivariate normal data, Spearman$_F$ and Spearman$_{BW}$ have been shown to be approximately accurate. The 95 % confidence interval usually has approximately .95 coverage, except with extreme values of ρ, in which case the coverage probability has dipped as low as .90 with Spearman$_F$ (Puth, Neuhäuser, & Ruxton, 2015; Rosner & Glynn, 2007). Given that the ranks are unaffected by monotonic transformations of the data, these bivariate normal results should generalize to simulated nonnormal data, and both Spearman methods should produce good coverage in most cases.

Nonlinear data transformations, such as the log transform, can sometimes convert nonnormal data to approximately normal data. In order to consider several such transformations simultaneously, it is useful to examine the performance of the Box–Cox transformation family (Box & Cox, 1964). This family has a free parameter, and depending on the parameter's value, the transformation can become equivalent to or an approximation of several other familiar transformations, including the logarithmic, square-root, and inverse transformations. Because of this flexibility, the Box–Cox transformation can approximately normalize a wide variety of skewed distributions, though it is less successful when applied to symmetric nonnormal distributions (e.g., extreme kurtosis but no skew). For the purpose of constructing a confidence interval of the correlation, the X and Y variables can be Box–Cox transformed in an attempt to approximately normalize them, and then the Fisher z' method can be used as usual.

An even more general approach to transformation can be found in RIN transformation, which is quite old, though somewhat obscure in psychology. The minor variations in this procedure have led to several names for the transformation, including normal scores (Fisher & Yates, 1938), Blom transformation (Blom, 1958), van der Waerden transformation (van der Waerden, 1952), and rankit transformation (Bliss, 1967; for a review, see Beasley, Erickson, & Allison, 2009). RIN transformation involves three steps. First, a variable is converted to ranks. Second, the ranks are converted to a 0-to-1 scale using a linear function. Finally, this distribution is transformed via the inverse of the normal cumulative distribution function (i.e., via probit transformation). The result is an approximately normal distribution regardless of the original shape of the data, so long as ties are infrequent and n is not

too small. For correlations with nonnormal data, RIN transformation has been shown to increase power (Bishara & Hittner, 2012; Puth et al., 2014), and also to reduce the bias and error of the point estimate (Bishara & Hittner, 2015). If RIN transformation is applied to the *X* and *Y* variables prior to use of the Fisher *z′* confidence interval, the confidence interval could likewise be more accurate; the marginal distributions will be approximately normal, though of course there is no guarantee that bivariate normality will be satisfied.

Bootstrap methods

In cases in which a linear relationship on the original scale is important, perhaps the most promising methods are those involving the bootstrap (Efron, 1979). Bootstrap methods involve resampling with replacement from the observed data, and so do not require assumptions about bivariate normality.

For correlations, a common approach is the nonparametric bivariate bootstrap (Lunneborg, 1985). In this approach, *n* pairs of data are randomly sampled with replacement (i.e., some rows of data might be sampled more than once). The Pearson correlation in this bootstrap sample is recorded. This procedure is repeated thousands of times, recording the new correlation each time. The distribution of the recorded correlations then provides an estimate of the sampling distribution of *r*. To construct the 95 % confidence interval, the 2.5th percentile of the recorded *r*s forms the lower bound, and the 97.5th percentile forms the upper bound. In addition to the unadjusted method described above, two adjusted measures are also considered here: an asymptotic adjustment (AA) and bias correction with acceleration (BCa). Previous research suggested that none of these approaches provides a perfect solution to the problem of nonnormality with correlated data. These methods can slightly inflate Type I error (Beasley et al., 2007; Bishara & Hittner, 2012; Lee & Rodgers, 1998; Rasmussen, 1987; Strube, 1988). Additionally, their 95 % confidence intervals can have actual coverage rates that range from 91 % to 99 % (Puth et al., 2014, 2015). However, if the data are extremely nonnormal and the population correlation is nonzero, bootstrap methods can at least be more accurate than the Fisher *z′* (e.g., Puth et al., 2014).

Given these inadequacies of the nonparametric bootstrap, it is important to consider a variation developed specifically for nonnormal data. Perhaps the most promising such variation is the *observed imposed bootstrap* (Beasley et al., 2007). In this bootstrap method, the sampling frame is created not from the *n* original pairs of *X* and *Y*, but rather from all possible pairs (n^2) of *X* and *Y*. The possible pairs are then rotated to recreate the originally observed *r* value, and then the bootstrap method proceeds as usual, with pairs sampled with replacement, the new *r* recorded, and so on. This approach allows for a much larger sampling frame, and hence a smoother bootstrap distribution, than the traditional nonparametric bootstrap. Beasley and colleagues (2007) examined this method for null hypothesis testing and showed that it could achieve good control over Type I error rates, particularly when data were nonnormal. A natural extension of their approach would be to use the observed imposed bootstrap distribution of *r* to generate confidence intervals. As with the nonparametric bootstrap, for the observed imposed bootstrap, we considered confidence intervals from unadjusted percentiles, AA, and BCa.

The present study

The purpose of the present study was to compare the accuracies of the various types of 95 % confidence intervals of the correlation in the context of nonnormal continuous data. There are several major differences between this study and previous work. First, we examined a wider array of confidence interval methods than previous studies have (e.g., Puth et al., 2014). Specifically, we examined 11 confidence interval methods: Fisher *z′*, two variants of Spearman confidence intervals, Box–Cox transformation, RIN transformation, three nonparametric bootstraps (unadjusted percentiles, AA, and BCa), and likewise, three observed imposed bootstraps. Such a comparison, along with the R code provided for each method (see Supplementary Materials A), could help researchers who are trying to choose and use such methods. Second, we examined the relative effectiveness of the observed imposed bootstrap for confidence intervals. This method has previously been examined only with hypothesis testing (Beasley et al., 2007), but because it was more accurate than other correlation bootstrap methods, it may also fare well with confidence intervals. Its performance relative to transformation methods is generally unknown. Third, we orthogonally manipulated skewness and kurtosis, at least to the extent that was mathematically possible, so that the effects of the two could be disentangled. Previous examinations of nonnormality with the correlation coefficient have compared different distributions (e.g., normal, chi-squared, etc.), but in doing so, changes in skewness were confounded with changes in kurtosis (e.g., Bishara & Hittner, 2012; Hayes, 1996; Kowalski, 1972; Puth et al., 2014). Finally, we examined the usefulness of sample information—sample skewness, kurtosis, and tests of sample nonnormality—in determining whether the default parametric method would be accurate. Sample information is important to consider because the shape of the population is often unknown to a researcher, and also because it is unclear how extremely nonnormal the sample must be in order to justify the use of alternative methods.

There is no known way of accomplishing this comparison through a formal proof, at least not for finite sample sizes, and so Monte Carlo simulations had to be used. The primary

dependent measure of interest was coverage probability, which should be approximately .95 if the 95 % confidence interval is accurate. Among confidence intervals that achieve this, a shorter confidence interval is more precise, and thus preferable to a longer one. To represent the wide array of nonnormality in actual data sets (Blanca et al., 2013; Micceri, 1989), our simulations involved a systematic manipulation of realistic values of population skewness and kurtosis. To assess the generality of the results, over 900 scenarios were examined.

Method

Scenario design

Overall, the factorial design consisted of 46 skewness/kurtosis combinations, two shape combinations, five sample sizes, and two population correlations. This design resulted in a total of 920 scenarios.

To define skewness and kurtosis, first, define the kth central moment in the population as

$$\mu_k = E\left[(x-\mu)^k\right], \tag{6}$$

where μ with no subscript is the population mean. The population skewness is

$$\gamma_1 = \frac{\mu_3}{\sigma^3}, \tag{7}$$

where σ is the population standard deviation. The population kurtosis is

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3. \tag{8}$$

Note that the population kurtosis is defined here with a constant of –3 so that, for a normal distribution, $\gamma_1 = \gamma_2 = 0$.

To inform the choice of our simulated skewness and kurtosis values, two published reviews of these values in actual data sets were considered. The first review covered 440 datasets in psychology and education with measures of knowledge, perception, and opinion (Micceri, 1989). The second review covered 693 datasets, mostly with measures of cognitive ability and personality (Blanca et al., 2013). Summarizing across these two reviews, estimated skewness values ranged from –2.5 to +2.3, and perhaps went even higher; Micceri categorized 17 % of datasets in a category that had skewness greater than 2, but he did not report the maximum skewness. Estimated kurtosis ranged from –1.9 to +37.4. To be on the safe side, we simulated skewness and kurtosis values slightly beyond these ranges, where possible.

The scenarios included nine values of population skewness (γ_1 = –4, –3, –2, –1, 0, 1, 2, 3, 4), and ten of population kurtosis (γ_2 = –1, 0, 2, 4, 6, 8, 10, 20, 30, 40). Figure 1a shows the specific combinations that were simulated. Many combinations of skewness and kurtosis were mathematically impossible, because the lower bound of kurtosis is determined by the squared skewness:

$$\gamma_2 \geq \gamma_1^2 - 2, \tag{9}$$

a boundary illustrated by the U-shaped curve in Fig. 1a. An additional constraint is that it becomes increasingly difficult to simulate nonnormal correlated data as the U-shaped boundary is approached (Headrick, 2010). Hence, 46 combinations of skewness and kurtosis were simulated, each represented by a letter or small square in Fig. 1a. Figure 1b–f show illustrative examples for select combinations of skewness and kurtosis.

We examined two types of distribution shape combinations: Either both X and Y had the same distribution shape, or only X was nonnormal and Y was normal. (Note that the bivariate normality assumption was satisfied here when both X and Y were normal. That is, not only were they marginally normal, but also all possible linear combinations of them were normal.) In order to understand the effect of sample size, we included five different samples sizes: n = 10, 20, 40, 80, and 160. To consider both zero and nonzero population correlations, two population correlation coefficients were used: ρ = 0 and .5.

Dependent measures

Coverage probability The observed coverage probability was the number of simulations in which a confidence interval covered the corresponding population parameter, divided by the total number of simulations. The population parameter of interest was Pearson's ρ for the Fisher z' and bootstrap methods. Raw data transformation methods required comparison to an appropriate population parameter for the respective transformation. For example, the Spearman confidence intervals should cover the population Spearman's ρ, which need not be equal to the population Pearson ρ that was set in the simulation. To estimate population parameters for the transformation approaches, within each scenario, a pseudo-population was generated with size N = 1,000,000. The estimated population parameter for Spearman confidence intervals was the rank-order correlation in this pseudo-population. The same strategy was taken with the Box–Cox and RIN transformations, for which the population parameter was estimated as the Pearson correlation of the pseudo-population following the Box–Cox and RIN transformations, respectively. For the interested reader, these parameters can be found in Supplementary Materials B.

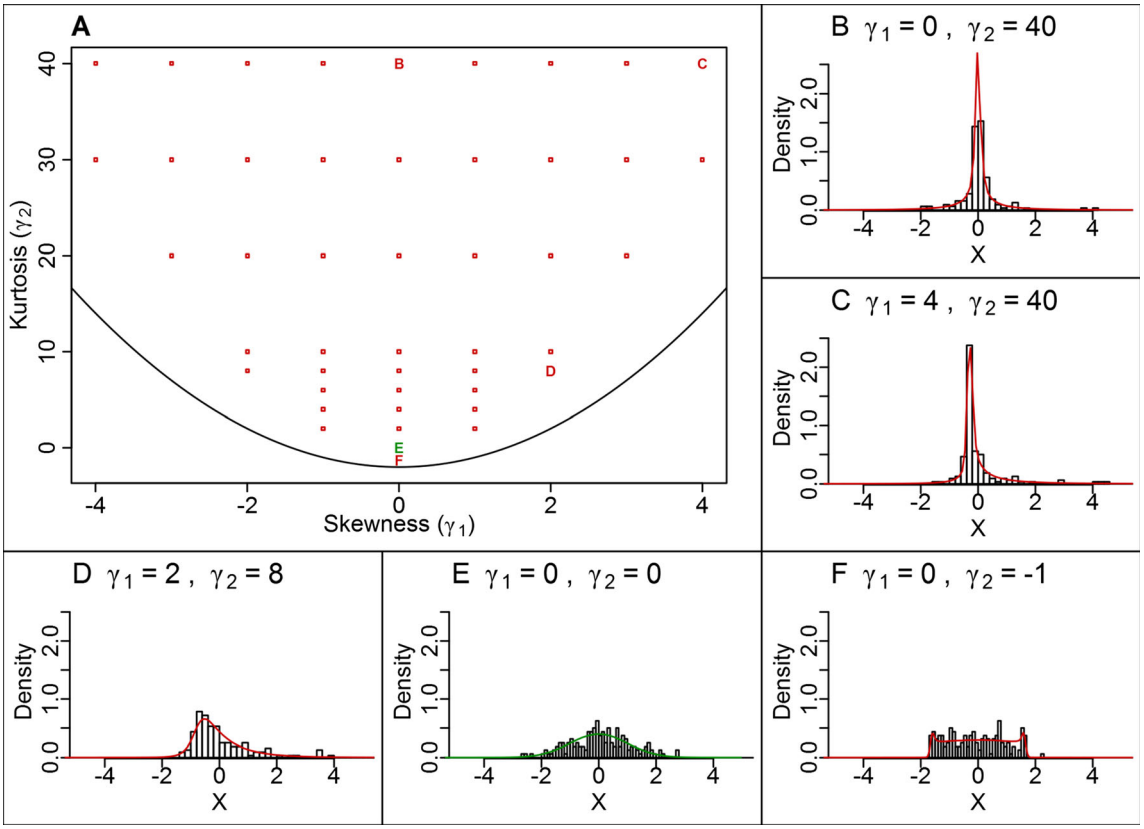

Fig. 1 **a** Letters and small squares represent combinations of population skewness and kurtosis values that were simulated, with red = nonnormal and green = normal. The U-shaped curve shows the lower boundary of kurtosis as a function of skewness (see Eq. 9). **b**–**f** The smaller panels show illustrative examples of select combinations of skewness and kurtosis. The smooth colored lines are approximate population densities, and the histogram bars illustrate a single random sample (n = 160) drawn from the population

Confidence interval length The confidence interval length was defined as the confidence interval's upper bound minus lower bound in a particular simulation.

Sample skewness and kurtosis Let the kth central moment of the sample be defined as

$$m_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k. \tag{10}$$

The unadjusted measures of sample skewness and kurtosis are typically defined as g_1 and g_2, respectively:

$$g_1 = \frac{m_3}{(m_2)^{3/2}}, \tag{11}$$

$$g_2 = \frac{m_4}{(m_2)^2} - 3. \tag{12}$$

When measured in small samples, the absolute skewness and kurtosis tend to be downward-biased. To mitigate this problem, g_1 and g_2 are adjusted on the basis of the sample size, resulting in G_1 (adjusted sample skewness) and G_2 (adjusted sample kurtosis):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1, \tag{13}$$

$$G_2 = \frac{n-1}{(n-2)(n-3)}\{(n+1)g_2 + 6\}. \tag{14}$$

In normal data, G_1 and G_2 are unbiased. In nonnormal data, they tend to be less biased than g_1 and g_2 (Joanes & Gill, 1998). Note that G_1 and G_2 are not the only possible adjusted measures, but they are perhaps the most popular and are used by default in several software packages, such as SPSS and Excel (Joanes & Gill, 1998). R code for G_1 and G_2 is provided in Supplementary Materials A, along with R code for other aspects of the method that may be useful to other researchers.

Tests of normality based on sample information To test nonnormality on the basis of the skewness or kurtosis in the sample, it is common to use the D'Agostino et al. (1990) tests. To test for significant skewness, g_1 is transformed to a z statistic, which is approximately normally distributed under the null hypothesis of population normality. For our purposes, in each simulation, a two-tailed alpha = .05 hypothesis test was

conducted on this z statistic (i.e., with cutoffs of approximately ±1.96). This is a univariate test, so it was done separately for the X and Y variables. Likewise, an additional set of two tests was done for kurtosis on the basis of g_2 in X and Y variables (for details, see D'Agostino et al., 1990). Note that D'Agostino et al. recommended their normal approximation test for kurtosis only if $n \geq 20$, though we found that it was still informative with $n = 10$ in the present simulations.

For an omnibus test of normality, we also examined the results of the Shapiro–Wilk test (Shapiro & Wilk, 1965). This test is also univariate, and must be done separately for X and Y. The Shapiro–Wilk test is often preferable to other tests (e.g., Kolmogorov–Smirnov) because it is usually more powerful (Shapiro, Wilk, & Chen, 1968).

Finally, we also examined multivariate tests of normality. Specifically, we examined Mardia's (1970, 1974) tests for multivariate skewness and multivariate kurtosis (see Eqs. 5.5 and 5.7 in Mardia, 1974).

Confidence interval construction methods

Fisher z See Eqs. 1 through 5.

Spearman rank-order with Fieller et al.'s (1957) standard error (Spearman$_{F}$) In this method, X and Y were separately transformed into ascending ranks. Then the Pearson correlation was computed on these ranks, thus forming the Spearman rank-order correlation coefficient. The confidence intervals were created just as in the Fisher z' method. However, the Fieller et al. estimate of the standard error was used, replacing the standard error in Eq. 4. Fieller et al.'s estimate of the standard error of z' is

$$\sigma_{z'_F} = 1.03 \Big/ \sqrt{N-3}. \tag{15}$$

Spearman rank-order with Bonett and Wright's (2000) standard error (Spearman$_{BW}$) Bonett and Wright's estimate of the standard error of z' is:

$$\sigma_{z'_{BW}} = \sqrt{1 + r^2 \Big/ 2} \Big/ \sqrt{N-3}. \tag{16}$$

All other details of this confidence interval method are the same as in the previous method.

Box–Cox transformation X and Y were separately transformed using the Box–Cox transformation (Box & Cox, 1964), which is especially well-suited for skewed data:

$$f(x,\lambda) = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda}, if\,\lambda \neq 0, \\ \ln(x), if\,\lambda = 0. \end{cases} \tag{17}$$

In each simulated sample, the free parameter λ was chosen by an iterative one-parameter optimization. The optimization's goal was to maximize normality, as measured by the correlation of the coordinates of the normal qq-plot, a correlation that tends to be higher with more normal distributions (Filliben, 1975). The optimization routine involved both golden-section search and successive parabolic interpolation, and was implemented through R's optimize function. Following transformation, the confidence interval was constructed via the Fisher z' method.

RIN transformation X and Y were separately transformed through the rankit formula (Bliss, 1967):

$$g(x) = \Phi^{-1}\left(\frac{x_r - 1/2}{n}\right), \tag{18}$$

where Φ^{-1} is the inverse of the cumulative normal distribution function, and x_r is the ascending rank of each x value. RIN transformation provides a good approximation of the unknown transformations that would normalize the unknown population distributions (Klaassen & Wellner, 1997; Zou & Hall, 2002). The rankit formula comes from a larger class of RIN transformations, all of which transform nearly any continuous distribution into an approximately normal one (see Beasley et al., 2009, for a review). The rankit formula was chosen because it more accurately reproduces the even moments of a normal distribution (Solomon & Sawilowsky, 2009). For a tutorial, see the Appendix of Bishara and Hittner (2012).

Nonparametric bootstrap For each simulation, a sample of n pairs of observations was drawn with replacement from the observed data. This bootstrap sample was then used to calculate the bootstrap correlation, r_1^*. The bootstrap correlation was calculated with the usual formula (Eq. 1), but with the bootstrap sample rather than the observed data. This process was then repeated for a total of 9,999 bootstrap samples, each with its own value of r^*. The 95 % confidence interval of the correlation was estimated at the 2.5th and 97.5th percentiles of the distribution of r^* (Efron, 1979; Lunneborg, 1985). In order to avoid undefined r^* values, any bootstrap sample that consisted entirely of repeats of a single pair was discarded and replaced (see Strube, 1988).

Nonparametric bootstrap with AA The percentile confidence interval bounds were widened by a factor of $\sqrt{(N+2)/(N+1)}$ (Efron, 1982). Specifically, if LB and UB are the original lower and upper percentile

bounds of the nonparametric bootstrap, then the AA confidence interval is

$$CI_{AA} = \frac{1}{2}(LB + UB) \pm \frac{1}{2}(UB{-}LB)\sqrt{(N+2)\Big/(N+1)}. \quad (19)$$

This method is less common in the recent literature, but it was included in our simulations because previous work had shown some promise for it with an observed imposed bootstrap (Beasley et al., 2007).

Nonparametric bootstrap with BCa This method adjusts the percentile bounds of the nonparametric bootstrap so as to improve the coverage of the interval, with the error of intended coverage approaching zero more rapidly in the limit as n approaches infinity. Extensive details can be found in Efron and Tibshirani (1993).

Code for this approach was adapted from Efron and Tibshirani's (1993) S package. On rare occasions, the original packaged code approximated the lower confidence interval boundary as being the 0th percentile. R interprets a "0" index as empty, which caused R to recycle the upper bound for both the lower and upper boundaries, resulting in a confidence interval with zero width, which could not possibly cover ρ. To avoid this problem, the code was adjusted so that the 0th percentile was replaced by the smallest r^* in the distribution of bootstrap replicates.

Observed imposed bootstrap In this method (Beasley et al., 2007), let $\{X', Y'\}$ be an initial sampling frame created by combining all possible pairs of the originally observed X and Y variables. For example, if $n = 3$, and $\{x_i,y_i\}$ represents the ith pair of the observed data, then the initial sampling frame will have $n^2 = 9$ pairs: $\{X', Y'\} = \{(x_1,y_1), (x_2,y_1), (x_3,y_1), (x_1,y_2), (x_2,y_2), (x_3,y_2), (x_1,y_3), (x_2,y_3), (x_3,y_3)\}$. This initial sampling frame will necessarily have a correlation of 0. Next, let the standardized sample frame be $\{X'', Y''\}$, where each x_j' has the mean of X' subtracted and is then divided by the standard deviation of X', and likewise for y_j', so as to standardize each variable. Next, to impose the originally observed correlation, r, the frame is transformed through bivariate Cholesky decomposition:

$$y_j'' = r*x_j' + \sqrt{1-r^2}*y_j', \quad (20)$$

where y_j''' is each new y-value (Kaiser & Dickman, 1962). Let $x_j'' = x_j'$. Importantly, the correlation between X''' and Y''' is the same as the original sample correlation, r. Thus, the "observed" correlation has been "imposed" on a much larger sampling frame that consists of n^2 instead of n pairs. From this final sampling frame of $\{X''', Y'''\}$, n pairs of observations are sampled with replacement, and the same procedure is followed as with the nonparametric bootstrap described above. Further details about the observed imposed bootstrap can be found in Beasley et al. (2007).

Observed imposed bootstrap with AA The same procedure was applied as with the nonparametric bootstrap, except that LB and UB were generated from the observed imposed bootstrap before application of Eq. 19.

Observed imposed bootstrap with BCa For this method, one change had to be made to Beasley et al.'s (2007) procedure, to reduce computing time. In pilot simulations with large samples, more than 90 % of the computing time for the entire simulation was devoted to calculating the observed imposed BCa confidence intervals. This computational burden occurred because of the jackknife estimation of the BCa's acceleration term, a jackknife estimation that results in looping across n^2 observations in this particular technique. To alleviate this computational burden in large samples, an approximate jackknife technique was used whenever n^2 exceeded 1,000. This approximation involved a random sample (without replacement) of 1,000 "leave-one-out" subsamples instead of all n^2 subsamples. This approximation only affected the acceleration term, not the number of bootstrap samples or the bias correction, and was only applied to scenarios with $n \geq 40$. The consequence could be a slightly less precise acceleration term in such situations. However, any resulting injury to the coverage rate of this method was not large enough to be noticeable when comparing across sample sizes or other observed imposed methods.

Simulation

Within each scenario, 10,000 simulations were conducted. This number was used so that the primary dependent variable—coverage probability—could be estimated with a 95 % confidence interval margin of error of less than ± .01. With this level of precision, any observed coverage probability less than .94 can be considered significantly below the ideal coverage probability of .95. Within each simulation, each bootstrap type (nonparametric and observed imposed) used a total of 9,999 bootstrap samples (Beasley & Rodgers, 2012; Boos, 2003).

In order to simulate nonnormal data with the specified population correlation, skewness, and kurtosis values, the fifth-order power polynomial method was used (Headrick, 2002; Headrick, Sheng, & Hodis, 2007). This method yields skewness and kurtosis values with more precision and less bias than do earlier, third-order power polynomial methods (Olvera Astivia & Zumbo, 2015). Further details of this data-generating method can be found in Supplementary Materials C. Simulations were conducted in the language R (R Development Core Team,

2014). Simulations were distributed on a high-performance computing cluster, with different cores devoted to different scenarios.

Results

Coverage probability and confidence interval length

Figure 2 gives a concise summary of the observed coverage probabilities of the different confidence interval estimation methods. An ideal 95 % confidence interval method would cover the true population correlation with probability .95, regardless of the scenario. Many methods reached this goal when the normality assumption was met, indicated by the circles in the figure. However, when considering the nonnormal scenarios (squares), most methods' coverage varied either higher or lower than .95. Of particular concern, the default method, Fisher z', led to some 95 % confidence intervals that covered the true parameter as little as 67.6 % of the time.

Among the transformation methods, the Spearman$_F$ and RIN approaches led to more desirable results, with coverage of approximately .95 regardless of the scenario. The Spearman$_{BW}$ method led to slightly higher than .95 coverage. The Box–Cox transformation method had low coverage in some scenarios, likely due to its inability to address symmetrical nonnormality ($\gamma_1 = 0$, $\gamma_2 \neq 0$).

Among the six bootstrapping methods, observed imposed AA and observed imposed BCa fared the best, with coverage of approximately .95 in the normal scenarios, and coverage always greater than .90 in the nonnormal scenarios. The observed imposed methods sometimes exceeded the target .95 coverage, suggesting that they might have been too long.

As is shown in Table 1, some confidence interval methods led to smaller (i.e., more precise) intervals than others. On average, RIN transformation led to the second smallest confidence intervals, bested only by the nonparametric bootstrap. Of course, the nonparametric bootstrap came with the cost of a lower-than-.95 coverage probability. The observed imposed bootstrap methods often had the longest intervals, particularly the observed imposed AA. Observed imposed BCa provided coverage probability similar to the AA method, but with more precise intervals.

Overall, among the transformation methods, the Spearman$_F$ and RIN intervals provided consistently accurate coverage probabilities, and the RIN interval lengths were especially precise. Among the bootstrap methods, which have the virtue of preserving the original scale of the raw data, there was no perfect alternative to the Fisher z'. At least, though, the observed imposed BCa method had an adequate coverage probability (>.90) and had more precise intervals than the corresponding AA method. In the remaining analyses, we focus on the Fisher z' method as compared to three promising alternatives: Spearman$_F$, RIN, and observed imposed BCa.

Population shape

As is shown at the top of Table 2, the Fisher z' confidence interval worked as intended when bivariate normality was satisfied. Even when just one variable was normal, the coverage was not noticeably impaired. However, when both variables were nonnormal, coverage fell below the nominal 95 %.

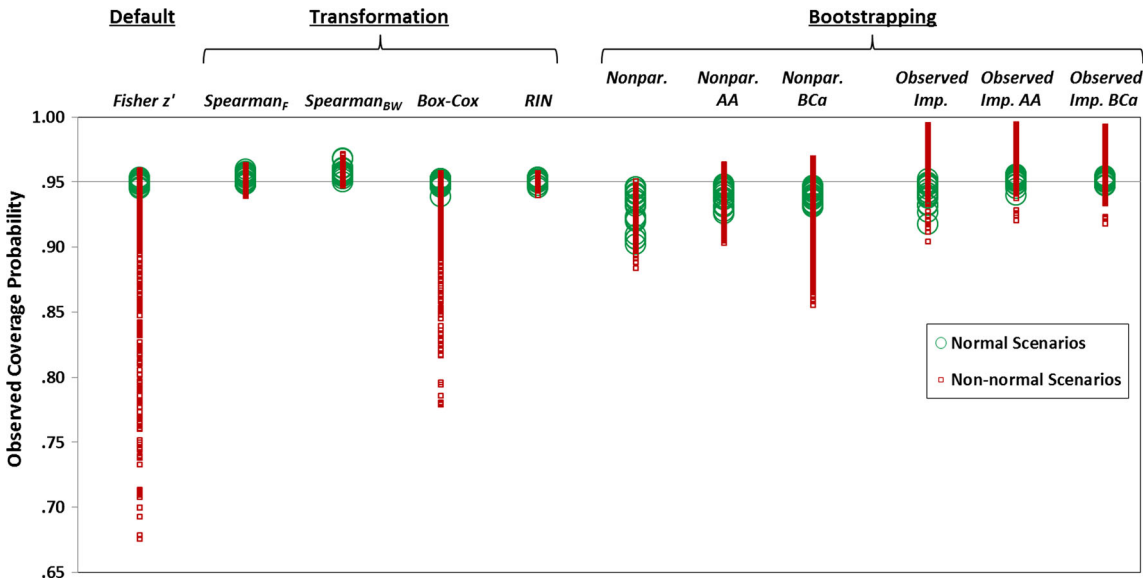

Fig. 2 A 95 % confidence interval will, ideally, cover the true population parameter with probability .95 in simulations. Observed coverage probabilities are shown for bivariate normal scenarios (circles) and for scenarios in which at least one variable was not normal (squares). Spearman$_F$ = Fieller et al. (1957), Spearman$_{BW}$ = Bonett and Wright (2000), RIN = rank-based inverse normal, Nonpar. = nonparametric, AA = asymptotic adjustment, BCa = bias correction and acceleration, Imp. = imposed

Table 1 Mean confidence interval length (upper bound minus lower bound)

Default Fisher z'	Transformation				Bootstrapping					
	$Spearman_F$	$Spearman_{BW}$	Box–Cox	RIN	Nonpar.	Nonpar. AA	Nonpar. BCa	Observed Imposed	Observed Imposed AA	Observed Imposed BCa
.597	.608	.613	.594	.590	.587	.635	.607	.651	.703	.663

Confidence interval lengths in this table have a margin of error no greater than ± .007, based on the 95 % confidence intervals of the means. RIN = rank-based inverse normal, Nonpar. = nonparametric, AA = asymptotic adjustment, BCa = bias correction and acceleration

The bold values in the table show proportions less than .940. This cutoff was chosen because the margin of error for the proportions was less than ± .010. All remaining tables also show proportions, and so also have margins of error less than ± .010 and cutoffs of .940.

The effect of sample size depended on the population correlation coefficient. With a zero correlation, larger sample sizes improved the coverage rate. However, with a nonzero correlation, larger sample sizes actually worsened the coverage rate.

As is shown in the lower panels of Table 2, the $Spearman_F$ and RIN methods produced approximately .95 coverage, regardless of population shape and sample size. The observed imposed BCa method showed approximately .95 coverage when bivariate normality was satisfied, but slightly higher coverage than intended when at least one variable was nonnormal.

As is shown in Table 3, both skewness and kurtosis affected the coverage rate of the Fisher z' method. The effect of negative skewness was approximately the same as that of positive skewness: Greater absolute skewness led to lower coverage. As can be seen by comparing the different rows, increased kurtosis also reduced coverage. It is difficult to compare skewness directly to kurtosis, but there are at least some circumstances in which the coverage probability appeared more sensitive to skewness than to kurtosis. For example, consider $\rho = .5$, $\gamma_1 = 0$, $\gamma_2 = 8$: Increasing skewness by 2 reduced coverage by .029. In contrast, increasing kurtosis by

Table 2 Observed coverage probabilities of 95 % confidence intervals among select methods as a function of sample size, population correlation (ρ), and population shape

Confidence Interval Method	n	$\rho = 0$			$\rho = .5$		
		Neither Normal	One Normal	Both Normal	Neither Normal	One Normal	Both Normal
Fisher z'	10	**.936**	.949	.950	**.893**	.950	.946
	20	**.940**	.950	.947	**.880**	.951	.951
	40	.944	.950	.951	**.862**	.952	.950
	80	.945	.950	.949	**.842**	.952	.949
	160	.947	.950	.951	**.823**	.950	.951
$Spearman_F$	10	.951	.951	.952	.950	.949	.949
	20	.954	.955	.955	.948	.948	.951
	40	.955	.956	.957	.947	.947	.948
	80	.956	.956	.956	.946	.946	.949
	160	.957	.957	.957	.946	.946	.949
RIN	10	.950	.950	.950	.948	.948	.947
	20	.949	.949	.948	.950	.950	.950
	40	.950	.949	.951	.950	.950	.950
	80	.950	.950	.950	.950	.950	.948
	160	.950	.951	.951	.949	.950	.952
Observed imposed bootstrap BCa	10	.970	.969	.953	.975	.971	.950
	20	.963	.963	.951	.971	.972	.948
	40	.962	.958	.954	.967	.975	.949
	80	.959	.954	.951	.964	.977	.949
	160	.956	.951	.951	.962	.980	.951

Bold values indicate coverage probabilities < .940. The different columns indicate how many raw variables (X and Y) were normally distributed in the population. RIN = rank-based inverse normal, BCa = bias correction and acceleration

Table 3 Observed coverage probabilities of Fisher z' confidence intervals as a function of population correlation, skewness, and kurtosis, when both variables were nonnormal

Population Correlation	Population Kurtosis (γ_2)	Population Skewness (γ_1)								
		–4	–3	–2	–1	0	1	2	3	4
$\rho = 0$	40	**.936**	**.934**	**.933**	**.932**	**.933**	**.932**	**.933**	**.935**	**.936**
	30	**.940**	**.939**	**.937**	**.937**	**.936**	**.938**	**.938**	**.939**	.943
	20		.945	.942	.941	.942	.940	.944	.946	
	10			.948	.946	.945	.945	.948		
	8			.950	.946	.947	.947	.948		
	6				.947	.946	.948			
	4				.947	.949	.949			
	2				.949	.948	.949			
	0					.950				
	–1					.949				
$\rho = .5$	40	**.750**	**.771**	**.774**	**.778**	**.781**	**.781**	**.779**	**.773**	**.753**
	30	**.771**	**.808**	**.818**	**.823**	**.823**	**.824**	**.818**	**.807**	**.774**
	20		**.837**	**.859**	**.868**	**.870**	**.868**	**.858**	**.835**	
	10			**.893**	**.911**	**.912**	**.911**	**.892**		
	8			**.892**	**.919**	**.922**	**.919**	**.893**		
	6				**.925**	**.931**	**.926**			
	4				**.932**	**.939**	**.930**			
	2				**.933**	.945	**.933**			
	0					.950				
	–1					.945				

Bold values indicate coverage probabilities < .940

the same amount reduced coverage by only .010. Of course, note that Table 3 omits scenarios in which a nonnormal *X* was paired with a normal *Y*, because the Fisher z' had approximately .95 coverage probability in such scenarios.

Sample shape

The results presented thus far rely on knowledge of the population skewness and kurtosis, which are rarely known to the researcher. From a practical perspective, it may be more important to know whether the choice to use Fisher z' can be informed by the observed sample. In other words, how much observed skewness, kurtosis, or other indication of nonnormality in the sample is sufficient to justify an alternative to the Fisher z'?

As is shown in Table 4, sample skewness and kurtosis could indeed be used to justify avoidance of the Fisher z'. As absolute sample skewness or sample kurtosis increased, the Fisher z' coverage probability decreased. Generally, the Fisher z' coverage probability was noticeably low when the absolute skewness was at least 1, or when kurtosis was at least 2. In contrast, the Spearman$_F$ and RIN coverage probabilities were approximately .95 regardless of the sample skewness or kurtosis, and observed imposed BCa coverage tended to exceed .95. Note that sample skewness and kurtosis are tightly restricted in small samples (see Cox, 2010). To keep the table estimates precise, only bins with at least 10,000 observations are shown.

As is shown in Table 5, normality tests of samples could also be used to justify avoidance of the Fisher z'. The D'Agostino et al. (1990) tests were most discriminating between accurate and inaccurate coverage situations. Specifically, for the Fisher z' interval, when both variables showed significant violations of skewness or kurtosis expected under normality, the coverage probability was noticeably poor. In contrast, when both variables showed nonsignificant results with this test, or if only one variable did, the coverage probability was approximately .95. The Shapiro–Wilk test was less discriminating between good and bad coverage situations, especially for small samples. Tests for multivariate skewness and kurtosis were least informative.

Discussion

Nonnormality can distort the Fisher z' confidence interval, and the outcome can be quite misleading; in the most extreme example, an intended 95 % confidence interval would have been better described as a "two-thirds" confidence interval.

Table 4 Observed coverage probabilities among select methods, based on sample size, sample skewness, and sample kurtosis

Confidence Interval Method	Sample Kurtosis (G_2)	n = 10				n = 40				n = 160			
		Absolute Sample Skewness ($\lvert G_1 \rvert$)											
		0–1	1–2	2–3	3 ≤	0–1	1–2	2–3	3 ≤	0–1	1–2	2–3	3 ≤
Fisher z'	20 ≤	–	–	–	–	–	–	–	–	–	–	–	**.704**
	10–20	–	–	–	–	–	–	**.759**	**.689**	**.837**	**.828**	**.807**	**.770**
	8–10	–	–	–	–	–	**.830**	**.795**	–	**.875**	**.865**	**.847**	–
	6–8	–	–	**.643**	–	**.869**	**.858**	**.829**	–	**.894**	**.883**	**.862**	–
	4–6	–	**.788**	**.722**	–	**.901**	**.884**	**.855**	–	**.915**	**.900**	–	–
	2–4	**.894**	**.866**	–	–	**.927**	**.907**	–	–	**.935**	**.919**	–	–
	1–2	**.930**	**.910**	–	–	**.940**	**.924**	–	–	.941	**.928**	–	–
	0–1	.946	**.928**	–	–	.947	**.932**	–	–	.946	–	–	–
	–1–0	.956	**.940**	–	–	.955	–	–	–	.954	–	–	–
	<–1	.959	–	–	–	.954	–	–	–	.949	–	–	–
Spearman$_F$	20 ≤	–	–	–	–	–	–	–	–	–	–	–	.950
	10–20	–	–	–	–	–	–	.953	.951	.950	.950	.950	.948
	8–10	–	–	–	–	–	.951	.952	–	.950	.950	.952	–
	6–8	–	–	.953	–	.951	.950	.951	–	.951	.951	.953	–
	4–6	–	.952	.955	–	.951	.951	.949	–	.951	.952	–	–
	2–4	.952	.952	–	–	.951	.951	–	–	.953	.952	–	–
	1–2	.951	.952	–	–	.951	.951	–	–	.951	.953	–	–
	0–1	.951	.951	–	–	.952	.954	–	–	.952	–	–	–
	–1–0	.950	.953	–	–	.951	–	–	–	.951	–	–	–
	<–1	.949	–	–	–	.952	–	–	–	.953	–	–	–
RIN	20 ≤	–	–	–	–	–	–	–	–	–	–	–	.950
	10–20	–	–	–	–	–	–	.952	.949	.950	.949	.949	.949
	8–10	–	–	–	–	–	.949	.951	–	.949	.950	.951	–
	6–8	–	–	.948	–	.948	.949	.950	–	.951	.949	.951	–
	4–6	–	.949	.952	–	.950	.950	.947	–	.950	.949	–	–
	2–4	.948	.949	–	–	.950	.950	–	–	.951	.950	–	–
	1–2	.949	.949	–	–	.951	.950	–	–	.950	.950	–	–
	0–1	.949	.949	–	–	.950	.951	–	–	.950	–	–	–
	–1–0	.949	.952	–	–	.950	–	–	–	.950	–	–	–
	<–1	.948	–	–	–	.951	–	–	–	.951	–	–	–
Observed Imposed BCa	20 ≤	–	–	–	–	–	–	–	–	–	–	–	.950
	10–20	–	–	–	–	–	–	.963	.960	.963	.961	.954	.940
	8–10	–	–	–	–	–	.965	.958	–	.964	.961	.949	–
	6–8	–	–	.955	–	.971	.968	.958	–	.965	.962	.949	–
	4–6	–	.974	.963	–	.972	.967	.955	–	.966	.960	–	–
	2–4	.986	.975	–	–	.973	.964	–	–	.965	.958	–	–
	1–2	.982	.981	–	–	.970	.962	–	–	.962	.955	–	–
	0–1	.978	.981	–	–	.966	.961	–	–	.964	–	–	–
	–1–0	.970	.982	–	–	.965	–	–	–	.966	–	–	–
	<–1	.959	–	–	–	.949	–	–	–	.941	–	–	–

Sample skewness and kurtosis ranges indicate the least amount of nonnormality observed in the X and Y variables. For example, the top number of the leftmost column, **.894**, represents the coverage probability when both $0 \leq \min(|G_{1X}|, |G_{1Y}|) < 1$ and $2 \leq \min(G_{2X}, G_{2Y}) < 4$. **Bold** values indicate coverage probabilities < .940. Only bins with at least 10,000 observations are reported. RIN = rank-based inverse normal, BCa = bias correction and acceleration

Table 5 Observed coverage probabilities among select methods as a function of sample normality test results

Confidence Interval Method	Test for Normality	$n = 10$			$n = 40$			$n = 160$		
		n.s.	One sig.	Both sig.	n.s.	One sig.	Both sig.	n.s.	One sig.	Both sig.
Fisher z'	Skewness–D'Agostino	.947	.947	**.796**	.947	.946	**.873**	.947	.944	**.875**
	Kurtosis–D'Agostino	.946	.944	**.761**	.949	.950	**.878**	.950	.950	**.883**
	Normality–Shapiro-Wilk	.947	.946	**.820**	.950	.951	**.885**	.952	.951	**.885**
	Multi. Skewness–Mardia	.942	**.898**		.948	**.918**		.949	**.913**	
	Multi. Kurtosis–Mardia	**.932**	–		.948	**.915**		.951	**.915**	
Spearman_F	Skewness–D'Agostino	.950	.951	.952	.951	.952	.951	.952	.951	.951
	Kurtosis–D'Agostino	.950	.951	.953	.952	.952	.951	.952	.952	.951
	Normality–Shapiro-Wilk	.950	.951	.952	.952	.952	.951	.954	.951	.951
	Multi. Skewness–Mardia	.953	.943		.954	.951		.953	.951	
	Multi. Kurtosis–Mardia	.950	–		.954	.950		.954	.951	
RIN	Skewness–D'Agostino	.948	.949	.949	.949	.950	.950	.951	.950	.950
	Kurtosis–D'Agostino	.948	.949	.950	.949	.950	.950	.950	.950	.950
	Normality–Shapiro-Wilk	.948	.949	.950	.950	.950	.950	.952	.950	.950
	Multi. Skewness–Mardia	.950	.943		.951	.949		.951	.950	
	Multi. Kurtosis–Mardia	.949	–		.950	.950		.951	.950	
Observed Imposed BCa	Skewness–D'Agostino	.959	.986	.972	.957	.971	.964	.956	.967	.959
	Kurtosis–D'Agostino	.962	.986	.969	.954	.970	.966	.952	.966	.960
	Normality–Shapiro-Wilk	.959	.984	.976	.954	.969	.966	.953	.966	.960
	Multi. Skewness–Mardia	.966	.987		.957	.969		.956	.963	
	Multi. Kurtosis–Mardia	.971	–		.957	.971		.952	.963	

Columns indicate whether neither, one, or both variables (X and Y) showed significant deviations from normality, $p < .05$. **Bold** indicates coverage probabilities < .940. Only bins with ≥10,000 observations are reported. RIN = rank-based inverse normal, BCa = bias correction and acceleration, Multi. = multivariate

Increasing sample size improved coverage when X and Y were independent, with a true population correlation of zero. However, increasing sample size worsened coverage with a nonzero population correlation. The latter pattern may appear counterintuitive, due to the misconception that the central limit theorem also applies to the sampling distribution of the Fisher z'. Unfortunately, the central limit theorem can provide no such comfort for correlations with nonnormal data. Thus, increasing sample size is not a general "cure-all" for nonnormality.

The present results show that Fisher z' coverage is affected not only by population kurtosis, but also by population skewness, and sometimes more so. Worse coverage could result from either high kurtosis *or* high absolute skewness, particularly when both variables were nonnormal. At least in some circumstances, Fisher z' coverage was somewhat more influenced by changes in skewness than kurtosis, which suggests that the historical emphasis on kurtosis may be misplaced, or at least incomplete.

Interestingly, poor Fisher z' coverage could be predicted by the corresponding sample statistics, which are sometimes the only values available to researchers. Sample statistics for higher-order moments, such as skewness and kurtosis, are sometimes considered untrustworthy due to their instability (Ratcliff, 1979). The present results suggest that, despite such concerns, sample statistics can provide clues as to the confidence interval coverage rate, and they can do so even in samples as small as 10. Significantly less than 95 % coverage occurred when both X and Y had an absolute sample skewness of 1 or higher, or when both had a sample kurtosis of 2 or higher. This pattern was in close agreement with the population pattern, in which similar thresholds for skewness and kurtosis emerged. Additionally, low coverage occurred if both X and Y showed significantly extreme skewness or kurtosis, based on the D'Agostino et al. (1990) tests. These results are important because they suggest that, even when researchers do not know the true shape of the population, they may be able to use sample skewness and kurtosis statistics, or hypothesis tests of those statistics, to help them decide whether to use the Fisher z' confidence interval.

If the default Fisher z' method cannot be used, a researcher's choice of alternatives may depend on whether raw data transformation is tolerable, or whether instead the measured correlation must be linear and on the original scales of X and Y. On the one hand, if raw data transformation is tolerable, at least two methods were quite accurate: the Spearman rank-

order correlation with Fieller et al. (1957) variance, and the RIN transformation. Both of these methods produced accurate coverage, and did so in all 920 scenarios, suggesting that they are robust to nonnormality. Of the two methods, RIN transformation led to slightly more precise (i.e., shorter) intervals.

On the other hand, if the correlation must indeed be linear and on the original scales, raw data transformation may be less desirable. In such a situation, there appears to be no perfect solution, at least not among the methods examined here. At best, the observed imposed bootstrap with BCa usually exceeded .95 coverage by producing intervals that were somewhat long. The observed imposed bootstrap has previously been shown to perform well for hypothesis testing for nonzero ρs (Beasley et al., 2007). Our results show that this method also holds promise for confidence intervals. This bootstrap method has the advantage of increasing the number of possible sampled observations, thus acting somewhat like a smoothing method. However, unlike many other conceivable smoothing methods for bivariate nonnormal data, the observed imposed method preserves the observed sample correlation, while also preserving the marginal density of one variable; the marginal density of the other variable is approximately preserved. Possible improvements to this method might involve further smoothing and/or preservation of both marginal densities, perhaps through iterative algorithms.

Limitations

In the present simulations, adequate Fisher z' coverage could be achieved if just one variable was normally distributed. However, this result is unlikely to hold in all situations, and especially so in large-n situations, as the Fisher z' sampling distribution converges toward the wrong value of variance (Hawkins, 1989). In "big data" research, inadequate Fisher z' coverage may occur with small deviations from normality in even one variable. Of course, with an extremely large n, the confidence intervals may become so narrow that they are generally no longer relevant.

In the simulations here, the Spearman$_F$ confidence intervals showed a slightly more accurate coverage rate than Spearman$_{BW}$ confidence intervals. This pattern does not appear to be universal (cf. Puth et al., 2015). Fortunately, the differences between the Spearman confidence intervals are usually trivial, so researchers' choice of a particular Spearman confidence interval is unlikely to affect the conclusions.

Datasets with frequent ties will not be well normalized by RIN or other raw data transformation approaches, because such methods do not break ties, and thus cannot erase the modes created by tied data. Datasets with frequent ties are typically addressed through concordance statistics (e.g., Goodman–Kruskal gamma, Kendall's tau, etc.; see Puth et al., 2015; Ruscio, 2008; Woods, 2007).

General recommendations

With nonnormal data, the typical methods for calculating the correlation coefficient can be far from optimal. On the basis of the present and recent work (e.g., Beasley et al., 2007; Bishara & Hittner, 2012, 2015; Puth et al., 2014, 2015), several conclusions can be reached about choosing among alternatives. On the one hand, if the sample size is small ($n < 20$) or if the linear correlation needs to be measured on the original scale, an optimal strategy may include some combination of resampling methods, such as a permutation test and bootstrapping. Such a strategy can minimize Type I errors, reduce bias, and as we showed here, provide a cautiously wide confidence interval, particularly with the observed imposed BCa bootstrap. On the other hand, if the sample size is at least moderate ($n \geq 20$) and there is no need to measure the linear correlation on the original scale, an optimal strategy may focus instead on the Spearman or RIN transformation methods. Both methods protect against Type I and II errors, reduce the bias and error of the point estimate, and provide approximately accurate confidence interval coverage. Of these two methods, RIN transformation often leads to slightly higher power and, relatedly, slightly more precise confidence intervals. R code for each of these methods can be found in Supplementary Materials A.

How extreme of a deviation from normality is needed to justify these alternative methods? The present study suggests two possible criteria. One criterion would be to use sample skewness and kurtosis estimates. That is, one could use alternative methods if both variables have absolute sample skewness greater than 1, or both have sample kurtosis greater than 2. The sample skewness and kurtosis equations used here are the same as those used in popular software packages such as SPSS, SAS, and Excel, and so such cutoffs can be easily implemented. A slightly more lenient criterion would be to rely on hypothesis tests of skewness and kurtosis, and only to use an alternative method if both variables show statistically significant violations of either univariate skewness or univariate kurtosis, as indicated by the D'Agostino et al. (1990) tests. R code for both sample estimates and the D'Agostino et al. tests can be found in Supplementary Materials A. Future simulation research may suggest slightly different sample nonnormality thresholds if simulations involve different ranges or proportions of population parameters. Given that the present correlation study was the first to orthogonally manipulate skewness and kurtosis, these cutoffs can be taken as starting points, and modified if necessary as further research indicates.

Finally, a word of caution should be noted. It is sometimes said that the Pearson correlation is "robust" to nonnormality (e.g., Havlicek & Peterson, 1977; for an early review, see Kowalski, 1972). However, the present and recent studies suggest a more nuanced viewpoint. On the one hand, the Pearson correlation can be robust in terms of the point

estimate and the Type I error rate, which usually converge on the correct values as *n* increases. On the other hand, the Pearson correlation is *not* generally robust in terms of confidence intervals or power, even in large samples. Thus, at least for the Pearson correlation, nonnormal data should be approached cautiously and with a careful consideration of alternative methods.

Author note We thank William Beasley, Bo Kai, Jiexiang Li, Oscar Olvera Astivia, John Ruscio, Graeme Ruxton, James Young, and the anonymous reviewers for helpful feedback on this project. We also thank Todd Headrick for sharing Mathematica code for the power polynomial methods. Finally, we thank Allan Strand, Chris Thackston, and the School of Sciences and Mathematics for help with and access to the Daito high-performance computing cluster.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington: Author.

Beasley, W. H., DeShea, L., Toothaker, L. E., Mendoza, J. L., Bard, D. E., & Rodgers, J. (2007). Bootstrapping to test for nonzero population correlation coefficients using univariate sampling. *Psychological Methods, 12,* 414–433. doi:10.1037/1082-989X.12.4.414

Beasley, T., Erickson, S., & Allison, D. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics, 39,* 580–595. doi:10.1007/s10519-009-9281-0

Beasley, W. H., & Rodgers, J. L. (2012). Bootstrapping and Monte Carlo methods. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 407–425). Washington: American Psychological Association.

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with non-normal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods, 17,* 399–417. doi:10.1037/a0028087

Bishara, A. J., & Hittner, J. B. (2015). Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement, 75,* 785–804.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9,* 78–84.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41.

Bliss, C. I. (1967). *Statistics in biology*. New York: McGraw-Hill.

Blom, G. (1958). *Statistical estimates and transformed beta-variables*. New York: Wiley.

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika, 65,* 23–28.

Boos, D. D. (2003). Introduction to the bootstrap world. *Statistical Science, 18,* 168–174.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26,* 211–252.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. doi:10.1037/0003-066X.49.12.997

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Quantitative Psychology and Measurement, 1,* 26. doi:10.3389/fpsyg.2010.00026

Cox, N. J. (2010). Speaking Stata: The limits of sample skewness and kurtosis. *Stata Journal, 10,* 482–495.

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

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

D'Agostino, R. B., Belanger, A., & D'Agostino, R. B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *American Statistician, 44,* 316–321.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2,* 292–307. doi:10.1037/1082-989X.2.3.292

Development Core Team, R. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. **Retrieved from www.R-project.org**

Duncan, G. T., & Layard, M. W. J. (1973). A Monte-Carlo study of asymptotically robust tests for correlation coefficients. *Biometrika, 60,* 551–558.

Edgell, S. E., & Noon, S. M. (1984). Effect of violation of normality on the *t* test of the correlation coefficient. *Psychological Bulletin, 95,* 576–583. doi:10.1037/0033-2909.95.3.576

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1–26.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *SIAM CBMS-NSF Monograph, 38*. doi:10.1137/1.9781611970319

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton: CRC Press.

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace *p* values. *Journal of Psychology, 217,* 27–37.

Field, A. (2000). *Discovering statistics using SPSS for Windows*. Thousand Oaks: Sage.

Fieller, E. C., Hartley, H. O., & Pearson, E. S. (1957). Tests for rank correlation coefficients: I. *Biometrika, 44,* 470–481.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics, 17,* 111–117.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10,* 507–521.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3–32.

Fisher, R. A., & Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. Oxford: Oliver & Boyd.

Fouladi, R. T., & Steiger, J. H. (2008). The Fisher transform of the Pearson product moment correlation coefficient and its square: Cumulants, moments, and applications. *Communications in Statistics—Simulation and Computation, 37,* 928–944.

Gayen, A. K. (1951). The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non-normal universes. *Biometrika, 38,* 219–247. doi:10.1093/biomet/38.1-2.219

Haldane, J. B. S. (1949). A note on non-normal correlation. *Biometrika, 36,* 467–468.

Havlicek, L., & Peterson, N. (1977). Effect of the violation of assumptions upon significance levels of the Pearson *r*. *Psychological Bulletin, 84,* 373–377.

Hawkins, D. L. (1989). Using *U* statistics to derive the asymptotic distribution of Fisher's *Z* statistic. *American Statistician, 43,* 235–237.

Hayes, A. (1996). Permutation test is not distribution-free: Testing H_0: $\rho = 0$. *Psychological Methods, 1,* 184–198.

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth: Harcourt Brace.

Headrick, T. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics and Data Analysis, 40,* 685–711.

Headrick, T. C. (2010). *Statistical simulation: Power method polynomials and other transformations*. Boca Raton: Chapman & Hall.

Headrick, T. C., Sheng, Y., & Hodis, F. A. (2007). Numerical computing and graphics for the power method transformation using Mathematica. *Journal of Statistical Software, 19,* 1–17.

Hoekstra, R., Johnson, A., & Kiers, H. A. L. (2012). Confidence intervals make a difference: Effects of presentation mode on inferential reasoning. *Educational and Psychological Measurement, 72,* 1039–1052.

Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Statistician, 47,* 183–189.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika, 27,* 179–182.

Klaassen, C. A. J., & Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favourable. *Bernoulli, 3,* 55–77.

Kowalski, C. J. (1972). On the effects of non-normality on the distribution of the sample product moment correlation coefficient. *Applied Statistics, 21,* 1–12.

Lee, W.-C., & Rodgers, J. L. (1998). Bootstrapping correlation coefficients using univariate and bivariate sampling. *Psychological Methods, 3,* 91–103. doi:10.1037/1082-989X.3.1.91

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science, 26,* 1827–1832. doi:10.1177/0956797615616374

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171. doi:10.1111/1467-8721.ep11512376

Lunneborg, C. E. (1985). Estimating the correlation coefficient: The bootstrap approach. *Psychological Bulletin, 98,* 209–215. doi:10.1037/0033-2909.98.1.209

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519–530.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyã, Series B, 36,* 115–128.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166. doi:10.1037/0033-2909.105.1.156

Olvera Astivia, O. L., & Zumbo, B. D. (2015). A cautionary note on the use of the Vale and Maurelli method to generate multivariate, nonnormal data for simulation purposes. *Educational and Psychological Measurement, 75,* 541–567.

Pagano, M., & Gauvreau, K. (2000). *Principles of biostatistics* (2nd ed.). Pacific Grove: Duxbury Press.

Psychological Science. (2014*). Submission guidelines*. Retrieved September 3, 2014, from www.psychologicalscience.org/index.php/publications/journals/psychological_science/ps-submissions

Psychonomic Society. (2012). *Psychonomic society guidelines on statistical issues*. Retrieved from www.psychonomic.org/statistical-guidelines

Puth, M., Neuhäuser, M., & Ruxton, G. D. (2014). Effective use of Pearson's product-moment correlation coefficient. *Animal Behaviour, 93,* 183–189.

Puth, M., Neuhäuser, M., & Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour, 102,* 77–84.

Rasmussen, J. (1987). Estimating correlation coefficients: Bootstrap and parametric approaches. *Psychological Bulletin, 101,* 136–139. doi:10.1037/0033-2909.101.1.136

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin, 86,* 446–461. doi:10.1037/0033-2909.86.3.446

Rosner, B. (1995). *Fundamentals of biostatistics* (4th ed.). Belmont: Wadsworth.

Rosner, B., & Glynn, R. J. (2007). Interval estimation for rank correlation coefficients based on the probit transformation with extension to measurement error correction of correlated ranked data. *Statistics in Medicine, 26,* 633–646.

Ruscio, J. (2008). Constructing confidence intervals for Spearman's rank correlation with ordinal data: A simulation study comparing analytic and bootstrap methods. *Journal of Modern Applied Statistical Methods, 7,* 416–434.

Shapiro, S. S., & Wilk, M. B. (1965). Analysis of variance test for normality (complete samples). *Biometrika, 52,* 591–611.

Shapiro, S. S., Wilk, M. B., & Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association, 63,* 1343–1372.

Solomon, S. R., & Sawilowsky, S. S. (2009). Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods, 8,* 448–462.

Strube, M. (1988). Bootstrap Type I error rates for the correlation coefficient: An examination of alternate procedures. *Psychological Bulletin, 104,* 290–292. doi:10.1037/0033-2909.104.2.290

Triola, M. (2010). *Elementary statistics* (11th ed.). Boston: Addison-Wesley/Pearson Education.

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

van der Waerden, B. L. (1952). Order tests for the two-sample problem and their power. *Indagationes Mathematicae, 14,* 453–458.

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. RIP. *American Statistician, 68,* 191–195.

Woods, C. M. (2007). Confidence intervals for gamma-family measures of ordinal association. *Psychological Methods, 12,* 185–204. doi:10.1080/00031305.2014.917055

Zimmerman, D., Zumbo, B., & Williams, R. (2003). Bias in estimation and hypothesis testing of correlation. *Psicológica, 24,* 133–158.

Zou, K. H., & Hall, W. J. (2002). On estimating a transformation correlation coefficient. *Journal of Applied Statistics, 29,* 745–760. doi:10.1080/02664760120098801