Article

Reducing Bias and Error in the Correlation Coefficient Due to Nonnormality

Educational and Psychological Measurement
2015, Vol. 75(5) 785–804

DOI: 10.1177/0013164414557639
epm.sagepub.com

Anthony J. Bishara[1] and James B. Hittner[1]

Abstract

It is more common for educational and psychological data to be nonnormal than to be approximately normal. This tendency may lead to bias and error in point estimates of the Pearson correlation coefficient. In a series of Monte Carlo simulations, the Pearson correlation was examined under conditions of normal and nonnormal data, and it was compared with its major alternatives, including the Spearman rank-order correlation, the bootstrap estimate, the Box–Cox transformation family, and a general normalizing transformation (i.e., rankit), as well as to various bias adjustments. Nonnormality caused the correlation coefficient to be inflated by up to +.14, particularly when the nonnormality involved heavy-tailed distributions. Traditional bias adjustments worsened this problem, further inflating the estimate. The Spearman and rankit correlations eliminated this inflation and provided conservative estimates. Rankit also minimized random error for most sample sizes, except for the smallest samples (n = 10), where bootstrapping was more effective. Overall, results justify the use of carefully chosen alternatives to the Pearson correlation when normality is violated.

Keywords

correlation, Pearson, Spearman, transformation, nonnormal, normality

In the social sciences, nonnormality is so common that it is arguably the ‘‘norm.’’ An analysis of several hundred psychometric and achievement data distributions in education and psychology found that 31% were extremely asymmetric, 29% had more than one peak, and 49% had at least one extremely heavy tail (Micceri, 1989).

[1]College of Charleston, Charleston, SC, USA

Corresponding Author:
Anthony J. Bishara, Department of Psychology, College of Charleston, 66 George St., Charleston, SC 29424, USA.
Email: BisharaA@cofc.edu

Nonnormality is also the rule, rather than the exception, in measures of cognitive ability and personality (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013), and in measures of reaction time (Logan, 1992; Van Zandt, 2000). This widespread prevalence of nonnormality raises a concern: Is our standard tool box of statistics robust to this reality? One of the most common statistics in social scientific research is the Pearson correlation coefficient. Of course, the Pearson correlation is used to examine simple bivariate relationships, but it and its unscaled equivalent—covariance—are also used in numerous multivariate techniques, including principal components analysis, canonical correlation analysis, and discriminant function analysis. If the Pearson correlation estimate is deflated or inflated by nonnormality, then this problem may generalize to a wide variety of statistical techniques. The purpose of the present report is (1) to determine when nonnormality distorts the point estimate of the Pearson correlation coefficient and (2) to systematically compare major alternatives to the Pearson correlation coefficient to determine if they can mitigate this problem.

It has long been known that, even for normal data, the correlation coefficient estimate has a small bias (Fisher, 1915). This bias in normal data is conservative, leading to underestimation of the true absolute correlation coefficient. Importantly, this normal data bias is a concern only for small samples; the absolute bias becomes negligible (less than .01) for a sample size greater than 20. To correct for this bias, Fisher (1915) and others (e.g., Olkin & Pratt, 1958) have recommended various approximate adjustments. The problem, though, is that each of these adjustments assumes bivariate normality, and could potentially cause more harm than good when nonnormality is present.

With nonnormal data, the traditional Pearson product–moment correlation may mischaracterize relationships in more noticeable ways. When using the Pearson correlation with nonnormal data, Type I and Type II error rates may be inflated (Bishara & Hittner, 2012; Blair & Lawson, 1982; Hayes, 1996). Additionally, transforming normal data into nonnormal data can reduce the correlation coefficient's absolute magnitude (Calkins, 1974; Lancaster, 1957; also see Dunlap, Burke, & Greer, 1995). Nonnormality could lead to two types of distortion in the point estimate of the correlation. First, the distortion could be systematic, leading to predictable tendencies to overestimate or underestimate the population parameter (i.e., bias). Second, the distortion could also be random, leading the estimates to vary in either direction (i.e., random error).

If the Pearson correlation is indeed distorted, the important question is, ''Compared to what?'' (Efron, 1988). In place of the Pearson correlation, there are several alternative techniques for addressing nonnormality, but it is not clear which, if any, would fare better. Major alternatives to the Pearson correlation include bootstrapping, the Spearman rank-order correlation, and other nonlinear data transformations.

Bootstrapping has attracted much attention in the primary literature (Efron, 1979; Lee & Rodgers, 1998; Rasmussen, 1987). The most common bootstrap for

correlations is the bivariate bootstrap technique (Rasmussen, 1987). In this technique, pairs of data values (X and Y) are sampled with replacement, and the Pearson correlation coefficient is computed for this new sample. This procedure is repeated many times to build a bootstrap distribution of correlation coefficients. The average of this bootstrap distribution serves as the bootstrap estimate of the correlation coefficient. Additionally, bootstrap distributions can also be used more generally for hypothesis testing and for confidence interval estimation (see W. H. Beasley & Rodgers, 2009), and bootstrap distributions have shown promise for addressing diverse challenges with correlations (e.g., Chan, 2009; Padilla & Veprinsky, 2012, 2014).

Whereas resampling techniques such as the bootstrap have been prominent in the primary literature, textbooks often recommend that nonnormality be addressed through the Spearman rank-order correlation or through other transformation approaches (e.g., Cohen, Cohen, West, & Aiken, 2003; Field, 2000; Gay, Mills, & Airasian, 2009; Triola, 2010). In the Spearman rank-order correlation, the data are ranked, and then the Pearson correlation coefficient is computed from the ranks. The Spearman approach can often be useful for nonnormal data, as it can increase power while maintaining a low Type I error rate (Fowler, 1987; Zimmerman & Zumbo, 1993). Like the Pearson correlation, the Spearman rank-order correlation also has a negative bias, at least for normal data (Kendall & Gibbons, 1990; Zimmerman, Zumbo, & Williams, 2003). Indeed, simulation studies have found greater negative bias for Spearman's approach versus Pearson's correlation (Arndt, Turvey, & Andreasen, 1999; Rosner & Glynn, 2007). Thus, the Spearman approach might be conservative.

The Spearman approach can be thought of as a member of a more general approach involving nonlinear transformation of the data prior to assessing the Pearson correlation. That is, the Spearman rank-order correlation simply involves a nonlinear transformation where the original distribution shapes are converted into uniform distributions (i.e., flat distributions of ranks, assuming there were no ties). Because the Spearman rank-order correlation transforms any initial distribution into a flat one, the shape of the original distribution should not matter. That is, the Spearman rank-order correlation should generally show a conservative bias for data that were originally normal or originally nonnormal.

Other transformation approaches, such as the logarithmic or square-root transformation, have been suggested for converting nonnormal data to normal data (e.g., Cohen et al., 2003). There are numerous such approaches, and it is often difficult to choose which approach to take in a principled, a priori fashion. A slightly more general transformation approach is the Box–Cox transformation family (Box & Cox, 1964), which subsumes logarithmic and square-root transformations. The Box–Cox family is useful for transforming skewed data into comparatively more normal data, but it is not able to address all forms of nonnormality, such as bimodal distributions, or symmetric but heavy-tailed distributions.

Theoretically, for any continuous population distribution, there should exist some unknown ideal transformation function that can convert the population distribution shape into an approximately normal one. It so happens that a good approximation of this ideal transformation function is a Rank-based Inverse Normal (RIN) transformation. RIN transformation involves a rank transformation, followed ultimately by use of the cumulative inverse normal function (Bliss, 1967; Blom, 1958; Fisher & Yates, 1938; Tukey, 1962; Van der Waerden, 1952; see T. Beasley, Erickson, & Allison, 2009, for a review). This old but obscure transformation approach can produce approximate normality in the sample regardless of the original distribution shape, so long as ties are rare and the sample size is reasonable. Though there are many transformations that can often reduce skew (e.g., logarithmic, Box–Cox), RIN transformation can also successfully transform multimodal or highly kurtotic distributions. When applied prior to assessment of a correlation, RIN transformation of sample variables is as efficient as if the ideal transformation functions for the populations were known and used (Klaassen & Wellner, 1997; Zou & Hall, 2002).

Previous work has shown that RIN transformation can minimize Type I and II error rates when testing the significance of a correlation (Bishara & Hittner, 2012). In simulations, RIN transformation was compared with other transformation techniques, including an optimized Box–Cox transformation. Additionally, these transformation approaches were compared with the simple Pearson product moment correlation, the Spearman rank-order correlation, and resampling approaches, such as the permutation test. RIN transformation was one of only a few approaches that consistently maintained Type I error rates at or below nominal alpha. Importantly, for nonnormal data, RIN transformation often showed increased statistical power. When the sample size was at least 20, RIN transformation resulted in higher power than did other common correlational techniques, including Spearman's rank-order correlation and various bootstrapping approaches. In contrast, the traditional Pearson correlation sometimes suffered from inflated Type I and II error rates. For smaller sample sizes ($n \leq 10$), the best alternative to the Pearson correlation was a resampling approach, in particular, the permutation test (for similar results, see Puth, Neuhäuser, & Ruxton, 2014). The permutation test only provides a test of the Null Hypothesis, not an estimate of the correlation, and so another resampling approach—a bootstrap estimate—may be more relevant here.

One previous report (Zimmerman et al., 2003) found some evidence of a slight positive bias (up to + .05) for the Pearson correlation coefficient under conditions of nonnormality. It is unclear, though, how general or extreme this problem could be. More important, previous work has compared major alternatives to the Pearson correlation for their ability to reduce Type I and II error rates (Bishara & Hittner, 2012; Puth et al., 2014), but it is unclear whether such alternatives can safely mitigate bias and error in the point estimate. The point estimate (r) is important because it indicates the direction and strength of a relationship, thus providing different information than a hypothesis test does. Furthermore, correlation point estimates are often highlighted in research via presentation in correlation tables, which display point estimates for all

possible pairs of variables. Despite the widespread reporting of the correlation point estimate, and the widespread commonality of nonnormal data, there have not been any large-scale comparisons of alternative point estimates that might be robust to nonnormality.

Using Monte Carlo simulation, we compared five statistical approaches: the traditional Pearson correlation, the bootstrap estimate, the correlation after RIN transformation, Box–Cox transformation, and Spearman rank-order transformation. If the literature on Type I and II error rates generalizes to bias and error in point estimates, then the RIN transformation may be effective at dealing with nonnormality for moderate to large samples, but for smaller samples, a resampling approach (such as the bootstrap) may be more effective. Additionally, for each approach, the unadjusted procedure was compared with two minor adjustments for the known bias in the correlation coefficient in normal data (Fisher, 1915; Olkin & Pratt, 1958). Because these adjustments were derived for normal data, they may be less successful at correcting for nonnormal data. To identify the generality and boundary conditions of the distorted correlations problem, we conducted simulations across 180 scenarios, examining various combinations of distribution shapes, sample sizes, and true population correlation coefficients.

Method

Statistical Approaches

Bias and error were measured using five statistical approaches: the Pearson Correlation Coefficient, the Bootstrap Estimate, the Pearson Correlation Coefficient following Box–Cox Transformation, following the Spearman Rank-Order Transformation, and following the RIN transformation.

Pearson Correlation Coefficient. This was the traditional statistic for the correlation coefficient:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{1}$$

Bootstrap Estimate. Within each simulation, the sample of n pairs of observations was drawn n times with replacement. Sampling was done at the level of pairs (i.e., a pair of X and Y observations would stay intact). This sample, called the bootstrap sample, was then used in the typical equation for the Pearson correlation coefficient (Equation 1) to calculate r_j^*. This procedure was repeated 9,999 times, each time using a new random sampling of the pairs of data, and resulting in a new value of r_j^*. The mean of all r_j^* was the bootstrap estimate of the correlation coefficient:

$$r_b = \sum_{j=1}^{9999} \frac{r_j^*}{9999} \tag{2}$$

The odd number of bootstrap samples was based on earlier code used for hypothesis testing where the number was intended to help create more precise alpha-levels (Bishara & Hittner, 2012; see Boos, 2003). More important, for the purposes of evaluating bias and error, this number of bootstrap samples is more than adequate, as 200 is sufficient for most point estimation purposes (see Efron, 1988; Efron & Tibshirani, 1993).

Box–Cox Transformation. The Box–Cox transformation family (Box & Cox, 1964) is a set of transformations that are primarily effective at reducing skew. The family includes the log-transformation, and approximations of both the inverse (i.e., $1/x$) transformation and square-root transformation. Thus, testing the Box–Cox family allows for a test of all of these commonly used transformations. The particular effect of the Box–Cox transform depends on the value of a free parameter, λ:

$$f(x,\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, if\ \lambda \neq 0 \\ \log(x), if\ \lambda = 0 \end{cases} \quad (3)$$

Within each simulation, the parameter λ was selected so as to maximize the normality of the resulting transformed distribution. To do so, a one-parameter optimization algorithm was used to search for the λ that maximized normality, as measured by the linearity of the normal qq-plot of the resulting distribution. Linearity was operationalized as the correlation of the coordinates on the qq-plot (see Filliben, 1975). This optimization was done based on the sample observed in the simulation. In other words, the algorithm's input did not include the population shape, a situation analogous to that of researchers when they must choose transformations for data from unknown populations. The optimization was done separately for X and Y variables, so the optimization could not capitalize on chance to (further) inflate the correlation. The search had the constraint $-5 < \lambda < 5$. This constraint was informed by pilot tests that showed that larger ranges provided no benefit.

The Box–Cox transformation requires the addition of a constant to all raw scores to remove negative values. Unfortunately, the choice of this arbitrary constant can influence the results (see Dougherty, Thomas, Brown, Chrabaszcz, & Tidwell, 2015, for a related example). Osborne (2010) recommended adding the absolute minimum score plus 1. A slightly greater value (1.00001) was added to ensure transformed values were all greater than 0, although this slight change should be inconsequential.

Spearman Correlation (i.e., Rank-Order Transformation). Variables are transformed into ranks:

$$g(x) = x_r \quad (4)$$

where $x_r = 1$ for the smallest x, $x_r = 2$ for the second smallest x, and so on. The traditional Pearson correlation coefficient formula (Equation 1) is used on the transformed X and Y variables.

RIN Transformation. In this approach, data were RIN transformed prior to assessing the Pearson correlation. The rankit equation (Bliss, 1967) was used because, as compared with other RIN equations, it more accurately produces the even moments of a normal distribution (Solomon & Sawilowsky, 2009). The rankit equation is

$$h(x) = \Phi^{-1}\left(\frac{g(x) - 1/2}{n}\right) \tag{5}$$

where Φ^{-1} is the inverse normal cumulative distribution function (also known as the probit function), and n is the sample size. Other RIN equations are nearly identical (Blom, 1958; Fisher & Yates, 1938; Tukey, 1962; Van der Waerden, 1952), and simply involve addition or subtraction of small constants in the numerator or denominator. Because of their similarity, RIN transformation equations are approximately linear transformations of one another, and so results from the rankit equation are likely to generalize to other RIN equations (Tukey, 1962; also see T. Beasley et al., 2009). Perhaps more important, just like the Spearman correlation, the correlation following RIN transformation is unaffected by addition or subtraction of a constant from the raw data, or for that matter, any other monotonic transformation of the raw data.

Bias Adjustments

Each statistical approach was examined in three ways: no bias adjustment, the Fisher Approximately Unbiased (FAU; Fisher, 1915) adjustment, and an adjustment by Olkin and Pratt (OP adjustment; Olkin & Pratt, 1958). The FAU adjustment applies a small adjustment to r, an adjustment that becomes smaller as n increases:

$$r_{FAU} = r(1 + \frac{1 - r^2}{2n}) \tag{6}$$

The OP adjustment applies a slightly larger adjustment to r:

$$r_{OP} = r(1 + \frac{1 - r^2}{2(n - 3)}) \tag{7}$$

Scenarios

As shown in Figure 1, the simulated distribution shapes were Normal, Bimodal, Slightly Skewed, Extremely Skewed, and Heavy-Tailed. Distribution parameters were chosen to allow for comparison of the present results with those involving Type I and II error rates for nonnormal correlated data (Bishara & Hittner, 2012). The Bimodal distribution was created by sampling with equal probability from one of two normal distributions whose means were 5 standard deviations apart. This creates bimodality that is easily visible (see Figure 1). The Slightly Skewed distribution was created through a Weibull distribution with shape = 1.5 and scale = 1. The Weibull

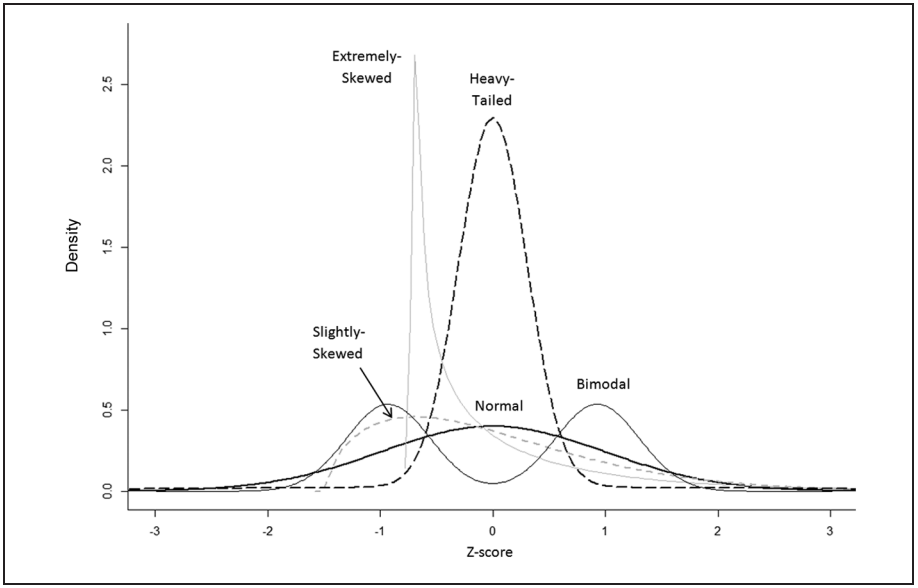

Figure 1. Normal and nonnormal distribution shapes used in simulations.

distribution is often used to model reaction times in a variety of tasks, including tasks that involve memory, perception, and reading (Berry, 1981; Logan, 1992). The parameters used here are near the bottom of the range of parameters common to human task performance times. They create a small amount of skew, usually the bare minimum amount of skew observed in reaction time data. To consider the other extreme of skew, the Extremely Skewed distribution was created via a χ^2 distribution with one degree of freedom. This Extremely Skewed shape represents approximately the uppermost amount of skewness found in Micceri's (1989) study of psychometric and achievement score datasets. A Heavy-Tailed distribution was examined to consider a case with high kurtosis but not skew. It was created by sampling from two normal distributions with the same mean but different standard deviations. Specifically, data were sampled with .9 probability from the first distribution, and .1 probability from a second distribution that had a standard deviation 10 times as large as the first. The skew and excess kurtosis of each distribution shape can be found in Table 1.

All possible scenarios were examined where X and Y had the same shape (e.g., both Bimodal) and also where X was normal but Y was not, leading to 9 combinations of distribution shapes. We examined different distribution shapes for X and Y since the uppermost limit of the Pearson correlation is reduced when under this condition (Nunnally & Bernstein, 1994).

The four true population Pearson correlation coefficients (ρ) were 0, .25, .50, and .75. The five sample sizes were n = 10, 20, 40, 80, and 160. At ns above 160, bias tends to be negligible (in our own simulations, when n was as high as 160, bias of r

Table 1. Descriptive Statistics for Distribution Shapes.

Shape	Skew	Excess kurtosis
Normal	0.0	0.0
Bimodal	0.0	−1.5
Slightly skewed	1.1	1.4
Extremely skewed	2.8	12.0
Heavy-tailed	0.0	22.3

Note. All shapes had population mean = 0, standard deviation = 1.

was no greater than +/−.01). Overall, a 9 (distribution shape) × 4 (true correlation) × 5 (sample size) factorial design resulted in 180 scenarios.

Simulation

Monte Carlo simulations were used to generate normal and nonnormal correlated data. Each scenario had 20,000 simulations. This number of simulations made the 95% confidence intervals no more than +/−.005 for bias estimates and +/−.004 for root mean squared error (RMSE) estimates within each scenario.

Correlated nonnormal variables were simulated with Ruscio and Kaczetow's (2008) algorithm. This algorithm is particularly flexible for creating mixture distributions, such as the Bimodal and Heavy-Tailed distributions used here. For each scenario, the algorithm was used to generate a population of 1,000,000 pairs of data with the target distribution shapes and population correlation coefficient, ρ. In each simulation, a sample of n pairs was drawn at random from this population. This approach allows for sampling error in each of the simulated data samples.

Code was written in R (R Core Team, 2014). Small sections of the code were adapted from Good (2009) for univariate nonnormal generation, and larger sections were adapted from Ruscio and Kaczetow (2008) for generating correlated data from these distributions. Large sections of code were also adapted from Bishara and Hittner (2012), primarily for transformations and bootstrap resampling.

Outcome Measures

The question of interest was this: How well does the observed statistic accurately represent the corresponding population parameter. The degree of inaccuracy was operationalized in two ways: Bias and RMSE. For example, within a particular scenario, bias of the Pearson Correlation was estimated as the mean of 20,000 sample Pearson correlations (r) minus the known population Pearson correlation (ρ) for that scenario. For the Spearman correlation, it was the mean sample rank-order correlation minus the known Spearman rank-order population correlation. Correlations for Box–Cox and RIN transformations were compared to their corresponding population

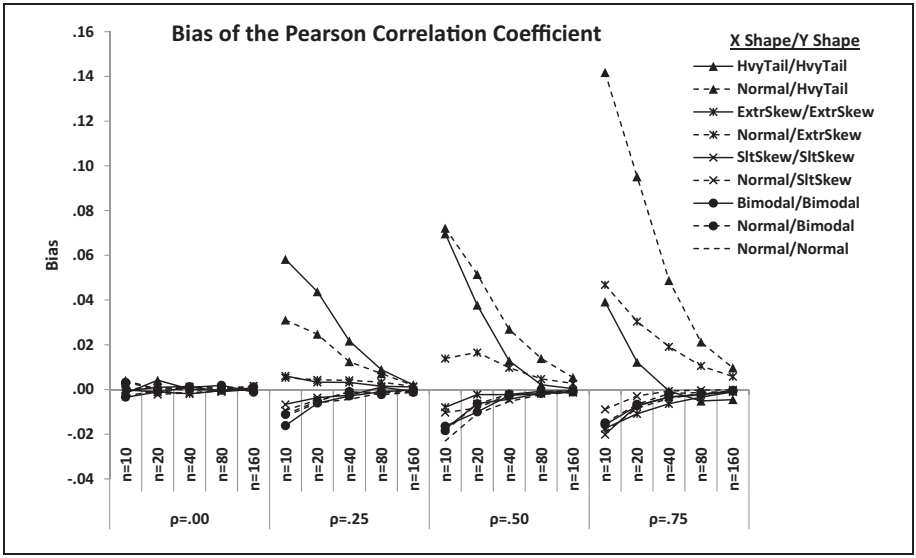

Figure 2. Bias of the Pearson *r* as a function of sample size (n = 10-160), true population correlation (ρ = 0-.75), and distribution shapes of the *X* and *Y* variables.
Note. The 95% confidence intervals of the mean for bias estimates were +/−.005 at most. HvyTail = Heavy-tailed; ExtrSkew = Extremely skewed; SltSkew = Slightly skewed.

correlations (i.e., following the same transformation of the population data). The one exception was the Bootstrap, where the mean bootstrap correlation was compared to the population Pearson correlation, which the bootstrap is intended to estimate.

RMSE was also defined relative to the corresponding population parameter. For example, RMSE of the Pearson correlation was defined as follows:

$$\mathrm{RMSE}_{\mathrm{Pearson}} = \sqrt{\frac{\sum_{k=1}^{20000} (r_k - \rho)^2}{20000}} \tag{8}$$

where r_k is the sample Pearson correlation of the kth simulation.

Results

Bias

As shown in Figure 2, the Pearson *r* could be exaggerated by nonnormal data. Bias could be as high as +.14, particularly with a Heavy-Tailed distribution for one variable and a small sample size (n = 10). Even with a more reasonable sample size (n = 40), bias could be as high as +.05. Positive bias was also noticeable if both distributions were Heavy-Tailed, or if one variable was Extremely Skewed. In contrast, the

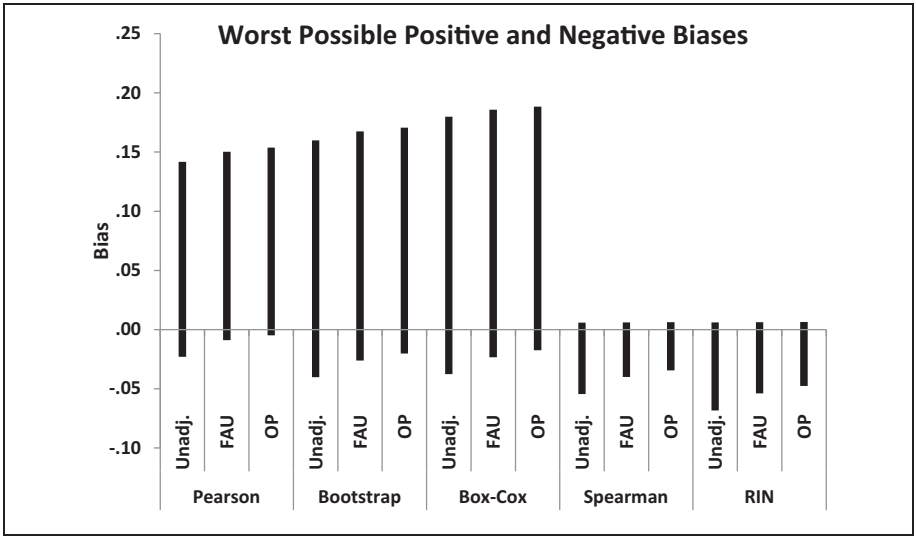

Figure 3. The range of bias across scenarios illustrates the worst possible positive and negative biases of various statistical approaches.
Note. Unadj. = unadjusted for bias; FAU = Fisher approximately unbiased adjustment; OP = Olkin and Pratt adjustment; RIN = rank-based inverse normal transformation.

absolute bias in the Normal/Normal case was much smaller, reaching no more than −.02. In general, bias was reduced as the sample size increased or if the true correlation coefficient (ρ) was small.

Some alternatives to the Pearson correlation more consistently corrected bias than did others. Figure 3 shows the range of bias, that is, the worst possible positive bias and the worst possible negative bias, for each alternative statistical approach and bias adjustment. The FAU and OP adjustments decreased the negative bias, but the decrease in negative bias came at the cost of an increase in positive bias. For example, the OP adjustment reduced the worst negative bias from −.023 to −.005, but increased the worst positive bias from +.142 to +.154. This pattern generally held true for the Pearson, Bootstrap, and Box–Cox approaches. However, this was not the case for Spearman rank-order and RIN estimates. The Spearman and RIN approaches never produced positive (exaggeration) bias above +.006. For those approaches, FAU and OP decreased the negative bias without a corresponding increase in the positive bias.

As shown in Figure 4, most approaches were susceptible to inflating the correlation coefficient when variables were Extremely Skewed or Heavy-Tailed. Only the Spearman and RIN approaches were generally immune to this problem, and showed a small but consistent negative bias. The OP adjustment helped mitigate this negative bias, leading to only minor negative biases for OP adjusted Spearman ($M = -.01$) and OP adjusted RIN ($M = -.01$). For the interested reader, bias estimates broken

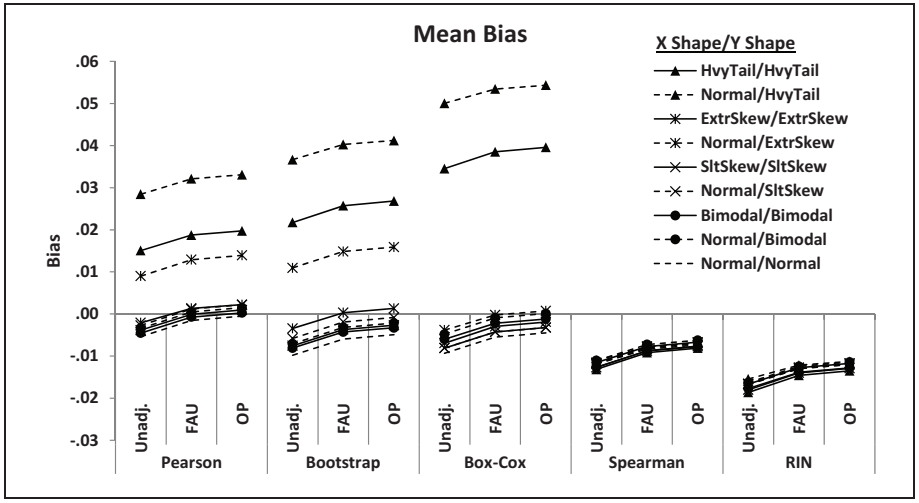

Figure 4. Mean bias as a function of statistical approaches and distribution shapes.
Note. HvyTail = Heavy-tailed; ExtrSkew = Extremely skewed; SltSkew = Slightly skewed; Unadj. = unadjusted for bias; FAU = Fisher approximately unbiased adjustment; OP = Olkin and Pratt adjustment; RIN = rank-based inverse normal transformation.

down by each of the 180 scenarios are available in online supplementary materials (via the first author's website).

Root Mean Squared Error

Whereas systematic distortions in the correlation estimate were measured with bias, random distortions were measured with RMSE. As shown in Figure 5, RMSE decreased as *n* increased (unsurprisingly). More important, the FAU and OP adjustments tended to increase error, and did so to a larger degree when *n* was small. This is because the equations for the adjustments lead to larger adjustments with small *n*. It is well-known that statistical corrections for the bias of an estimator often come at the cost of increased error of that estimator. Because the FAU and OP bias adjustments tended to increase error, further analyses of RMSE focus on unadjusted statistics.

As shown in Figure 6, the advantages of different approaches depended on the sample size. With small sample sizes, bootstrapping led to the smallest (best) RMSE. However, by an *n* of 20, RIN produced the best RMSE on average, and the advantages of RIN over bootstrapping became larger as *n* increased. As shown in Figure 7, most of RIN's benefit came from the extremely kurtotic distributions, particularly with an *n* of at least 20. At higher *n*s, RIN did produce a slight advantage even with more mild types of nonnormality (e.g., bimodality), but the advantage with such mild versions of nonnormality was very slight. Finally, as shown in Figure 8, RIN's

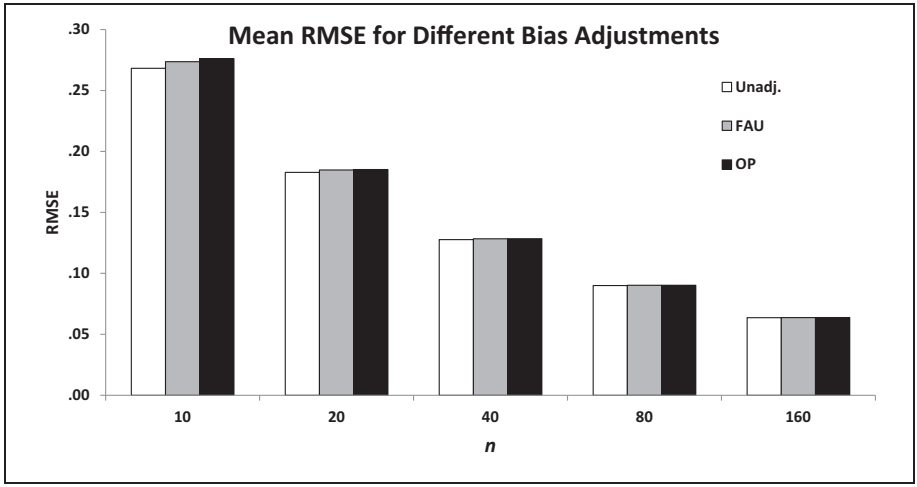

Figure 5. Mean RMSE as a function of bias adjustment and sample size.
Note. The 95% confidence intervals of the mean for RMSE estimates were +/−.004 at most. RMSE = root mean squared error; Unadj. = unadjusted for bias; FAU = Fisher approximately unbiased adjustment; OP = Olkin and Pratt adjustment.

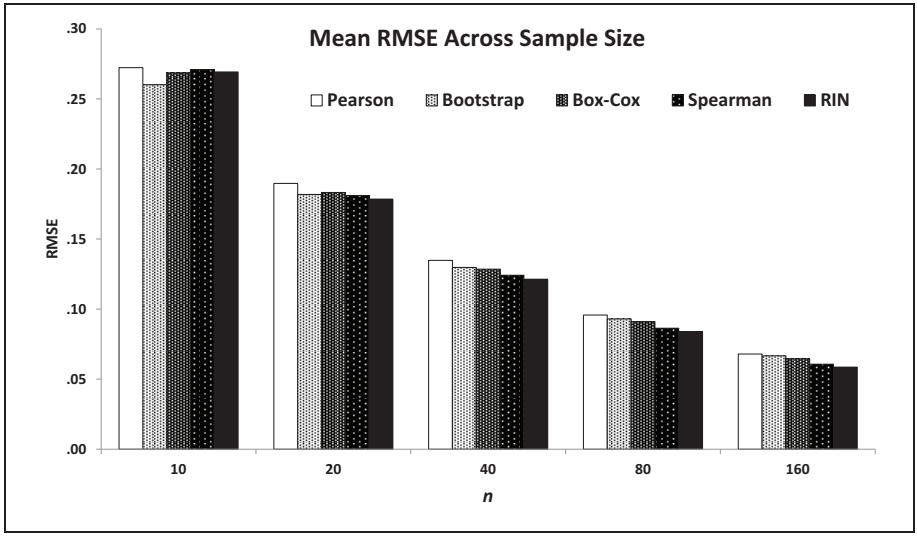

Figure 6. Mean RMSE as a function of statistical approach and sample size.
Note. RMSE = root mean squared error; RIN = Rank-based inverse normal transformation.

benefits for RMSE were more apparent at larger population correlation coefficients. More detailed RMSE results can be found in online supplementary materials.

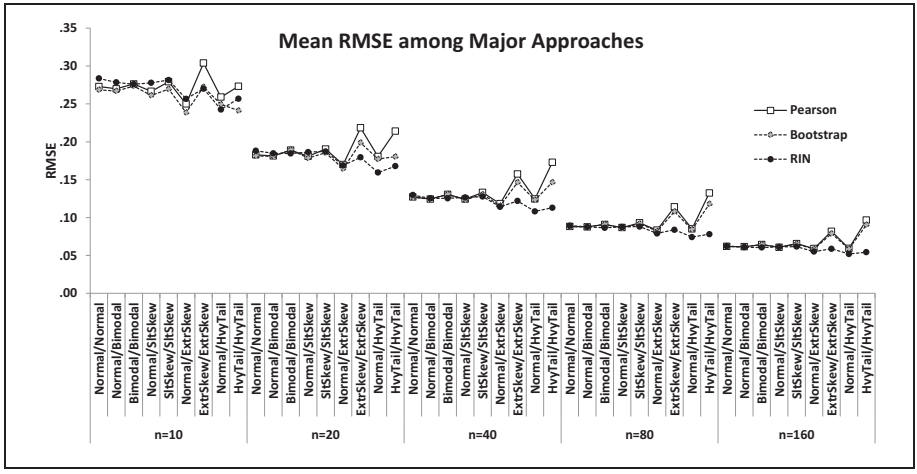

Figure 7. Mean RMSE among major approaches as a function of sample size and distribution shape.

Note. RMSE = root mean squared error; RIN = rank-based inverse normal transformation; HvyTail = Heavy-tailed; ExtrSkew = Extremely skewed; SltSkew = Slightly skewed.

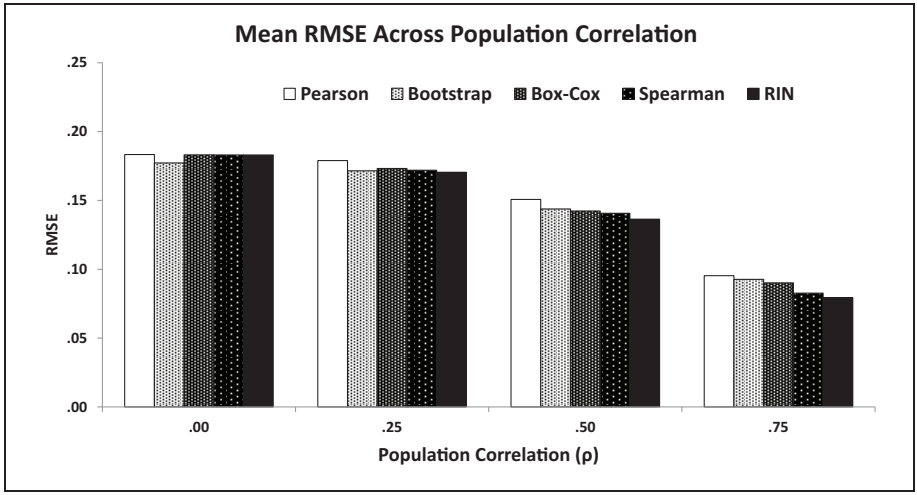

Figure 8. Mean RMSE as a function of statistical approach and true population correlation (ρ).

Note. RMSE = root mean squared error; RIN = rank-based inverse normal transformation.

Discussion

Historically, with the Pearson correlation, researchers have been concerned mainly with its slightly conservative bias. However, as shown in the current study, this bias is trivial when compared with the potential for liberal bias due to nonnormality. In

the Pearson correlation, the largest overestimation bias was approximately seven times the size of the largest underestimation bias. That is, nonnormality can lead to inflated estimates of the correlation coefficient, not to mention estimates that are susceptible to random error, as well.

Inflated correlations are possible when data distributions are extremely nonnormal, particularly with excess kurtosis, as in the Extremely Skewed and Heavy-Tailed distributions examined here. Such excess kurtosis distributions are prone to outliers at one or both extremes, respectively. Inflated correlations do not require outliers to be caused by contamination or measurement error; the problem occurred here even though the outliers were part of the population distribution and the underlying correlation of interest. Perhaps more worrisome is the possibility that the nonnormality most conducive to inflated correlations might go unnoticed. The Heavy-Tailed distribution (see Figure 1) appears symmetrical and bell-shaped. A casual glance at a histogram of a heavy-tailed sample might appear approximately normal. That is, the potential for exaggerated correlations could be easily missed.

Some alternatives to the Pearson correlation reduced such bias better than others did. The typical corrections for bias—FAU adjustment and OP adjustment—made exaggeration biases even worse. However, the Spearman and RIN correlations eliminated the exaggeration, providing conservative estimates with slightly negative biases. Overall, there was no fool-proof way of eliminating bias, but with extremely nonnormal distributions, Spearman and RIN correlations were safer in that they avoided exaggeration.

When attempting to reduce random distortions (RMSE), bootstrapping was effective for some situations and RIN for others. Bootstrapping was most effective when sample sizes were small or when the population correlation was small. However, with larger sample sizes and population correlations, the benefits of RIN transformation became apparent. On average, even with a sample size of just 20, the RIN transformation reduced error more effectively than bootstrapping. Most of this benefit came from the high kurtosis distributions (Extremely Skewed and Heavy-Tailed). For more modest normality violations (e.g., Bimodal), RIN transformation provided just a slight reduction in error, and only then at sample sizes of approximately 80 or more. RIN transformation may be less effective with small samples because the shape of a small sample distribution can poorly represent the shape of the population distribution. That is, the RIN transformation function of a small sample might be quite different from the ideal transformation function of the population from which that sample was drawn. Overall, these results are broadly consistent with the literature on hypothesis testing of correlations with nonnormal data: Moderate to large samples benefit from RIN transformation, but smaller samples are better aided by a resampling approach (Bishara & Hittner, 2012; Puth et al., 2014). Additionally, these results converge with others by suggesting that inferences based on the Pearson correlation are distorted primarily by high kurtosis (W. H. Beasley et al., 2007; Edgell & Noon, 1984; Hayes, 1996).

Results of RIN transformation were very similar to those of the Spearman rank-order transformation. This similarity is likely due to their similar methods. The Spearman rank-order correlation involves transforming the data into a flat distribution of ranks, whereas the RIN approach involves the additional step of transforming that flat distribution of ranks into an approximately normal distribution. Overall, the Spearman correlation was slightly better at reducing bias, but the RIN approach's bias was still conservative, and the RIN approach was more effective at reducing error.

Transformation may mitigate the effect of nonnormality on the Pearson correlation, but transformation is not always appropriate. Nonlinear transformations will not break ties, including those that occur with extreme ceiling or floor effects, and so even RIN transformation cannot approximately normalize data in such situations. For situations with large numbers of ties, it may be more appropriate to use concordance measures of association, such as the Goodman–Kruskal gamma or Kendall's tau (see Woods, 2007). A broader concern is that transformation (including transformation involved in the Spearman rank-order correlation) changes the nature of the relationship being measured, causing the resulting correlation coefficient to indicate the monotonic rather than the linear relationship. This can be problematic but only when two conditions are satisfied: Both variables have at least interval scales, and the theory of interest predicts a linear relationship. In psychology and education, it is rare for both of these conditions to be satisfied. Most measures require the creation and selection of several test questions, questions which have slightly different sensitivities to the underlying construct of interest. Thus, for example, the difference between a score of 10 and 11 on such a test is rarely identical to the difference between 20 and 21. That is, the necessary arbitrariness of the test construction can lead to arbitrariness in the scale of measurement. Perhaps more important, it is rare for the substantive theories of interest to be specific enough to predict a linear relationship rather than a monotonic relationship. The common practice of examining only the Pearson correlation relies on an implicit simplifying assumption that the relationship is linear, even though the true relationship might be monotonic but not strictly linear. Overall, transformation may be useful when ties are rare, and either at least one variable is not a true interval scale or linearity is not required by the theory of interest. In other situations, the traditional Pearson correlation and the bootstrap estimate have the advantage of allowing a linear interpretation, but they come with the cost of susceptibility to bias, as shown in our results.

The current simulations involved a wide array of scenarios, but it is impossible to examine every conceivable one. In the current simulations, nonnormality was present in the true populations. This type of nonnormality, if anything, probably underestimates the potential for distorted correlations. Nonnormality due to contamination or measurement error could further distort estimates. Additionally, in the current simulations, population correlations were all 0 or positive. Negative population correlations typically lead to mirror images of the bias patterns observed in positive correlations (Shieh, 2010). In other words, situations that produce exaggerated positive

correlations for positive effects are likely to produce exaggerated negative correlations for negative effects.

Overall, there are several conclusions that can be drawn from this research. First, nonnormality can cause exaggerated or otherwise distorted point estimates of the Pearson correlation coefficient, and particularly so when the nonnormality involves high kurtosis (heavy-tails). Second, some alternatives mitigate these problems better than others, and the choice of alternatives should depend on the sample size and distribution shape, as well as the relative importance of reducing bias versus random error. Finally, our results suggest caution more broadly when analyzing nonnormal data. Because correlations (and likewise covariances) underlie numerous statistical procedures, the distortions caused by nonnormality that we observed under bivariate conditions could also plague numerous multivariate procedures, leading to distorted point estimates more generally. This is not to suggest that the solutions offered here (e.g., RIN transformation and/or resampling approaches) will universally apply to all such situations, but at the very least, it does suggest that normality violations should not be ignored.

Authors' Note

Supplementary materials can be found on the first author's website: http://bisharaa.people.cofc.edu/.

Acknowledgments

We thank Bo Kai and Tal Yarkoni for helpful feedback on this project. We also thank Clayton McCauley and Allan Strand for help with the Department of Computer Science's high performance computing cluster.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Arndt, S., Turvey, C., & Andreasen, N. C. (1999). Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's tau correlation. *Journal of Psychiatric Research*, *33*, 97-104.

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., 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, W. H., & Rodgers, J. L. (2009). Resampling methods. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 362-386). London, England: SAGE.

Berry, G. L. (1981). The Weibull distribution as a human performance descriptor. *IEEE Transactions on Systems, Man, & Cybernetics*, *11*, 501-504. doi:10.1109/TSMC.1981.4308727

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

Blair, R., & Lawson, S. (1982). Another look at the robustness of the product-moment correlation coefficient to population non-normality. *Florida Journal of Educational Research*, *24*, 11-15.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(2), 78-84.

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

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

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 (Methodological)*, *26*, 211-252.

Calkins, D. S. (1974). Some effects of non-normal distribution shape on the magnitude of the Pearson product moment correlation coefficient. *Interamerican Journal of Psychology*, *8*, 261-288.

Chan, W. (2009). Bootstrap standard error and confidence intervals for the difference between two squared multiple correlation coefficients. *Educational and Psychological Measurement*, *69*, 566-584. doi:10.1177/0013164408324466

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Dougherty, M. R., Thomas, R. P., Brown, R., Chrabaszcz, J. S., & Tidwell, J. W. (2015). An introduction to the general monotone model with application to two problematic datasets. *Sociological Methodology*. Advance online publication. doi:10.1177/0081175014562589

Dunlap, W., Burke, M., & Greer, T. (1995). The effect of skew on the magnitude of product-moment correlations. *Journal of General Psychology*, *122*, 365-377.

Edgell, S., & Noon, S. (1984). Effect of violation of normality on the t test of the correlation coefficient. *Psychological Bulletin*, *95*, 576-583.

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

Efron, B. (1988). Bootstrap confidence intervals: Good or bad? *Psychological Bulletin*, *104*, 293-296.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

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

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

Fowler, R. (1987). Power and robustness in product-moment correlation. *Applied Psychological Measurement*, *11*, 419-428.

Gay, L., Mills, G., & Airasian, P. (2009). *Educational research: Competencies for analysis and applications* (9th ed.). Upper Saddle River, NJ: Merrill/Pearson.

Good, P. (2009). Robustness of Pearson correlation. *Interstat*, *15*(5), 1-6.

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

Kendall, M., & Gibbons, J.D. (1990). *Rank correlation methods* (5th ed.). New York, NY: Oxford University Press.

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.

Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika*, *44*, 289-292. doi:10.1093/biomet/44.1-2.289.

Lee, W., & Rodgers, J. L. (1998). Bootstrapping correlation coefficients using univariate and bivariate sampling. *Psychological Methods*, *3*(1), 91-103.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 883-914. doi:10.1037/0278-7393.18.5.883

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.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, *29*, 201-211.

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, & Evaluation*, *15*(12), 1-9.

Padilla, M. A., & Veprinsky, A. (2012). Correlation attenuation due to measurement error: A new approach using the bootstrap procedure. *Educational and Psychological Measurement*, *72*, 827-846. doi:10.1177/0013164412443963

Padilla, M. A., & Veprinsky, A. (2014). Bootstrapped deattenuated correlation: Nonnormal distributions. *Educational and Psychological Measurement*, *74*, 823-830. doi: 10.1177/0013164414531780

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

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

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

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., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research*, *43*, 335-381.

Shieh, G. (2010). Estimation of the simple correlation coefficient. *Behavior Research Methods*, *42*, 906-917.

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.

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

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, *33*, 1-67. doi:10.1214/aoms/1177704711.

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

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review*, *7*, 424-465.

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

Zimmerman, D., & Zumbo, B. (1993). Significance testing of correlation using scores, ranks, and modified ranks. *Educational and Psychological Measurement*, *53*, 897-904.

Zimmerman, D., Zumbo, B., & Williams, R. (2003). Bias in estimation and hypothesis testing of correlation. *Psicológica*, *24*(1), 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