METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Analytic and Bootstrap Confidence Intervals for the Common-Language Effect Size Estimate

Johnson Ching-Hong Li[a], Virginia Man Chung Tze[b]

[a] *Lab for Research in Quantitative and Applied Statistical Psychology (LIQAS), Department of Psychology, University of Manitoba, Manitoba, Canada.* **[b]** *Department of Educational Administration, Foundations, and Psychology, University of Manitoba, Manitoba, Canada.*

Methodology, 2021, Vol. 17(1), 1–21, https://doi.org/10.5964/meth.4495

Received: 2019-11-25 • **Accepted:** 2021-02-22 • **Published (VoR):** 2021-03-31

Corresponding Author: Johnson Ching-Hong Li, Department of Psychology, University of Manitoba, P508, Duff Roblin Building, R3T 2N2, MB, Canada. Phone: (1)-204-318-2923, E-mail: Johnson.Li@umanitoba.ca

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

Evaluating how an effect-size estimate performs between two continuous variables based on the common-language effect size (CLES) has received increasing attention. While Blomqvist (1950; https://doi.org/10.1214/aoms/1177729754) developed a parametric estimator (q') for the CLES, there has been limited progress in further refining CLES. This study: a) extends Blomqvist's work by providing a mathematical foundation for Bp (a non-parametric version of CLES) and an analytic approach for estimating its standard error; and b) evaluates the performance of the analytic and bootstrap confidence intervals (CIs) for Bp. The simulation shows that the bootstrap bias-corrected-and-accelerated interval (BCaI) has the best protected Type 1 error rate with a slight compromise in Power, whereas the analytic-t CI has the highest overall Power but with a Type 1 error slightly larger than the nominal value. This study also uses a real-world data-set to demonstrate the applicability of the CLES in measuring the relationship between age and sexual compulsivity.

Keywords

common-language effect size, confidence intervals, bootstrapping, Monte Carlo simulation, probability-of-superiority

In psychological research, there has been increasing attention paid to the importance of effect size (ES) estimates and confidence interval (CI) in improving the quality of statistical practices. Cumming (2014) provided detailed guidelines for researchers in reporting ES and CI when conducting meta-analysis, which is standard statistical practice in the

21st century. Many psychological associations (e.g., American Psychological Association, 2010) also state that researchers should report ES and CI, because it is considered the best reporting strategy for research studies.

ES is mainly measured and quantified based on two theoretical frameworks: the *d*-family (group difference) and *r*-family (correlation). The idea of *d* comes from the *standardized mean difference between two groups of observations* (e.g., gender difference on performance), whereas the idea of *r* is based on the level of *linear association between two variables* (e.g., correlation between cognitive ability and performance). One benefit of measuring and presenting ES is that the strength of a study effect can be measured and disseminated in an understandable, interpretable, and replicable manner (May, 2004). For example, a *d* of .20 suggests that female job incumbents outperform male job incumbents by .20 *SD* units in an appraisal test. Cohen (1988) also provided guidelines for *d* in behavioral science, and he concluded that levels of *d* equal to .20, .50, and .80 were commonly found in behavioral research, which corresponds to a small, medium, and large ES, respectively.

By contrast, the interpretation of *r* is much more challenging (Brooks, Dalal, & Nolan, 2014). Cohen (1988) attempted to provide an interpretation of *r*. Researchers can take a square of *r* (e.g., $r^2 = .09$; known as the *proportion of variance explained*) in order to interpret the proportion of variance of variable *Y* (e.g., performance) that can be accounted for by variable *X* (e.g., cognitive ability). For example, $r = .30$ can be interpreted as 9% of variance of incumbents' performance can be accounted for by their cognitive ability in an organization. Cohen (1988) provided a general rule for interpreting a small ($r = .10$), medium ($r = .30$), and large ($r = .50$) ES. Despite Cohen's efforts in providing such an interpretation, r^2 remains a challenging concept. First, it is hard for researchers and practitioners to truly understand the meaning of *proportion of variance explained* without first fully comprehending the meaning of the variance of a variable (*Y*), and how this variable can be overlapped with or explained by the variance of another variable (*X*). Second, the criteria for a small, medium, and large ES are $r^2 = .01$ (or 1% variance of *Y* is explained by *X*), .09, and .25, respectively; this can seem confusing and arbitrary to researchers and practitioners. Even some students and researchers in psychology may not be comfortable with this kind of statistical terminology (e.g., Brooks et al., 2014).

In light of this, researchers have explored and considered alternative ESs beyond the *d*-family and *r*-family. On the basis of the probability-of-bivariate-superiority (PBS) theory, researchers (e.g., Cliff, 1993, 1994, 1996; Cliff & Keats, 2003; McGraw & Wong, 1992; Li, 2016, 2018a; Li & Waisman, 2019; Ruscio, 2008) proposed the idea of common-language effect size (CLES), which is regarded as a more understandable and interpretable ES than *r* and *d*. For example, instead of saying that there is a *d* (standardized mean difference) of 1.00 on a cognitive ability test between the treatment group and control group, a researcher can express that there is a 76% likelihood (CLES $= \Phi\left(d/\sqrt{2}\right)$, where Φ is the normal cumulative distribution function, and data are assumed to follow normal

distribution; Ruscio, 2008) that a randomly-selected treatment group participant will perform better on a cognitive ability test than a randomly-selected control group participant.

While CLES for two group comparisons has received increasing attention among behavioral researchers, the understanding of whether CLES can be used in evaluating the effect between two continuous variables X and Y is limited. Dunlap (1994) is one of the pioneer studies that seeks to fill in this research gap. Assuming that X and Y are continuous variables that follow a bivariate linear and normal correlation (BLNC), researchers can obtain Pearson's correlation r and convert it to a CLES, which is labelled as CL_r in Dunlap's study. That is,

$$CL_r = sin^{-1}(r)/\pi + .5 \quad (1)$$

where CL_r is the CLES that explains the effect between X and Y, r is Pearson's correlation coefficient, sin^{-1} is the inverse sine function, and π is a constant ($\approx$ 3.14159). For example, instead of saying that 16% (r = .4, or r^2 = .16) of variance of sons' heights is explained by variance in their fathers' heights, one can state that "a father who is above average in height has a 63% likelihood of having a son of above-average height" (Dunlap, 1994, p. 510). The mathematical proof for Dunlap's r-to-CL_r conversion is shown in Li and Waisman (2019): if there is a linear correlation between normally distributed X and Y continuous scores, then the x-plane and y-plane can be divided into four quadrants based on the lines $x = \bar{x}$ and $y = \bar{y}$ (where $\bar{x}$ is the mean of X and $\bar{y}$ is the mean of Y). An observed correlation between X and Y (r) can then be converted to its corresponding CL_r through Equation 1 on the basis of the number of sample observations (n_1) that belong to the first or third quadrants compared with the total number of observations (n).

Despite the potential of CL_r, its use is relatively limited in practice because a) researchers may perceive that Dunlap's (1994) CL_r is merely a r-translated statistic useful for better knowledge mobilization only, and b) there is no analytic method for estimating the standard error (*SE*) and CI for this estimate. Indeed, Li and Waisman's (2019) study shows that the bivariate linear and normal correlation (BLNC) conditions are not necessary for researchers to obtain and interpret PBS. Instead, researchers can use and report the non-parametric version of CL_r, which is known as B_p in Li and Waisman. This study aims to extend Blomqvist's (1950), and Li and Waisman's (2019) work by proposing and developing analytic methods (i.e., analytic-z and analytic-t) for estimating the *SE* and CI surrounding B_p, which offers the necessary mathematical foundation for B_p. This can be used by both theoretical researchers who are interested in further testing and generalizing B_p to other data scenarios (e.g., multivariate relationships) and by applied researchers who are interested in evaluating their data based on B_p, and comparing the performance of these methods with the empirical methods (i.e., bootstrap percentile interval [BPI], bootstrap bias-corrected-and-accelerated interval [BCaI], bootstrap stand-

ard interval [BSI] based on the empirical z distribution [BSI-z], and BSI based on the empirical t distribution [BSI-t] in Li & Waisman, 2019).

Review of Li and Waisman's (2019) PBS

Blomqvist (1950) developed a likelihood-based statistic (q). Assuming that (x_i, y_i), where $i = 1,2,\ldots, n$ be n samples from a two-dimensional population associated with a BLNC-based cumulative distribution function (cdf),

$$f(x,y) = e^{-\frac{1}{2(1-r^2)}\left[\left(\frac{x-\bar{x}}{s_x}\right)^2 - 2r\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) + \left(\frac{y-\bar{y}}{s_y}\right)^2\right]} / 2\pi s_1 s_2\sqrt{1-r^2} \tag{2}$$

where r is the sample correlation, $\bar{x}$ is the sample mean of x, $\bar{y}$ is the sample mean of y, s_x is the sample SD of X, and s_y is the sample SD of Y. Blomqvist's Equation 12 proved that r can be mathematically linked to q, on the basis of the number of sample observations (n_1) that belong to the first or third quadrants compared with the number of sample observations (n_2) that belong to the second or fourth quadrants in a x-y plane. That is,

$$q \equiv \frac{2}{\pi} sin^{-1}(r) \tag{3}$$

where "≡" is the equal sign, when the condition of BLNC is met.

The PBS between X and Y does not necessarily depend upon the BLNC data condition assumed in Equation 2. Rather, a randomly selected point (x, y) is assumed to fall into 1 of the 4 quadrants (a, b, c, and d) in a x-y plane,

$$f(x,y)\begin{cases} a(x_i, y_i), \text{if } P(x_i > \bar{x} \wedge y_i > \bar{y}) \\ b(x_i, y_i), \text{if } P(x_i \le \bar{x} \wedge y_i > \bar{y}) \\ c(x_i, y_i), \text{if } P(x_i \le \bar{x} \wedge y_i \le \bar{y}) \\ d(x_i, y_i), \text{if } P(x_i > \bar{x} \wedge y_i \le \bar{y}) \end{cases} \tag{4}$$

Given Equation (4), n_1 is defined as the number of $a(x_i, y_i)$ and $c(x_i, y_i)$ points, n_2 is defined as the number of $b(x_i, y_i)$ and $d(x_i, y_i)$ points, and $\wedge$ is the logical function of "and". With the additional conditions—a) the population means (or medians) are uniquely defined as a certain point (e.g., 0); and b) x and y never equals to 0 because of a continuous distribution—Blomqvist proved another estimator (called q'; Equation 1 in Blomqvist's study) that estimates the aforementioned likelihood parameter q. That is, q is estimated by q' through

$$q' = \frac{n_1 - n_2}{n_1 + n_2} = \frac{2n_1}{n_1 + n_2} - 1 \tag{5}$$

where n_1 is the number of observations that belong to the first or third quadrants, and n_2 is the number of observations that belong to the second or fourth quadrants in the x-y plane. Li and Waisman (2019) provided a proof between PBS (B_p) and Blomqvist's likelihood estimate (q) in order to show how Dunlap's (1994) CL_r is only applicable when data meets the assumption of BLNC. First, given Equation 4, Li and Waisman formally defined PBS as

$$\begin{aligned} B_p &= P(Y_i > \bar{Y} \wedge X_i > \bar{X}) = n_1/(n_1 + n_2) \\ &= \sum_{i=1}^{n} \#\left[sign(x_i - \bar{x}) \cdot sign(y_i - \bar{y}) > 0\right]/(n_1 + n_2) \end{aligned} \tag{6}$$

where B_p is the sample PBS value, $X_i > \bar{X}$ denotes whether a X score of participant i is above the mean of all other X scores, and $Y_i > \bar{Y}$ denotes whether a Y score of participant i is above the mean of all other Y scores. Computationally, B_p can be effectively estimated through $\sum_{i=1}^{n} \#\left[\text{sign}(x_i - \bar{x}) \cdot \text{sign}(y_i - \bar{y}) > 0\right]/(n_1 + n_2)$.

Under the special case when X and Y follow BLNC, one can divide Equation 3 by 2 and add 0.5 to become

$$q'\left(\frac{1}{2}\right) + 0.5 \equiv \frac{1}{\pi} sin^{-1}(r) + 0.5 \tag{7}$$

where the left side becomes $n_1/(n_1 + n_2)$ [given $\left(\frac{2n_1}{n_1 + n_2} - 1\right)\left(\frac{1}{2}\right) + 0.5$ from Equation 5], such that

$$n_1/(n_1 + n_2) \equiv \frac{1}{\pi} sin^{-1}(r) + 0.5 \tag{8}$$

In fact, Equation 8 is identical to Dunlap's (1994) r-to-CL_r in Equation 1. This implies the algorithm of "$\frac{1}{\pi} sin^{-1}(r) + 0.5$" can be used for converting r to CL_r to measure PBS, if and only if BLNC is met ("≡"). Li and Waisman's (2019) simulation results showed that researchers can routinely use B_p in Equation 6 that is robust to data generated from either Equation 2 (BLNC) or Equation 4 (PBS).

The Proposed Analytical Methods for B_p

Li and Waisman's (2019) B_p provides a nonparametric method for obtaining a point estimate of PBS between two continuous variables. However, this method is not sufficient in practice. Researchers and practitioners have to evaluate and interpret the CI for B_p in order to evaluate the associated sampling error, precision, and significance. Moreover, the CI offers a range of possible B_p estimates for researchers to examine and replicate in reproducibility research (Cumming & Maillardet, 2006). One approach for obtaining the CI is using non-parametric bootstrapping, a computer-intensive technique that resamples data-sets with replacement many times (e.g., 2,000) in order to simulate the sampling

distribution for the 2,000 resampled B_p estimates and obtain the bootstrap-based CIs such as BSI-*z*, BSI-*t*, BPI, and BCaI. Li and Waisman found some good coverage probabilities of the true population B_p value (β_p) from the bootstrap CIs. A second approach is using an analytic method for estimating the *SE* and CI for B_p. The mathematical proof for deriving the *SE* for a new statistical measure is often sophisticated. Fortunately, this study proposes and demonstrates that one can use Blomqvist's (1950) proof for the *SE* of *q* (Blomqvist's likelihood estimate) and convert it to the *SE* of B_p in practice, as discussed below.

Assuming that (*x*, *y*) points are generated from a PBS-based function in Equation 4, Blomqvist (1950; Section 3, Equations 3 - 9) asymptotically derived the sampling distribution of the likelihood estimate (q'): specifically, when $n \to \infty$, q' is asymptotically and approximately distributed as a normal distribution, with an expected mean $E(q') \sim Q$ and *SE* equals to $\sigma(q') \sim \sqrt{(1-Q^2)/n}$, where Q is the true population likelihood value in Blomqvist. In practice, Q can be substituted by q'. Given that $q'\left(\frac{1}{2}\right) + 0.5 = n_1/(n_1+n_2) = B_p$, and the variance properties [$Var(a \cdot X) = a^2 Var(X)$ and $Var(a+X) = Var(X)$, where a is a constant], B_p is asymptotically distributed as a normal distribution with

$$E\left(B_p\right) = \beta_p \tag{9}$$

$$\sigma\left(B_p\right) \sim \sigma\left(q'\left(\frac{1}{2}\right) + 0.5\right) = \sigma\left(q'\left(\frac{1}{2}\right)\right) = \sigma\left(\frac{1}{4}\sqrt{\frac{1-q'^2}{n}}\right) = \sigma\left(\frac{1}{4}\sqrt{\frac{1-\left(\frac{n_1-n_2}{n_1+n_2}\right)^2}{n_1+n_2}}\right) \tag{10}$$

where β_p is the true population PBS value of B_p, and n_1 and n_2 are defined in Equation 5. Given Equation 10, the $(1-\alpha)\cdot 100\%$ (e.g., $(1-\alpha)\cdot 100\% = 95\%$, where α is the level of significance) analytic-*z*, symmetrical CI surrounding β_p can be constructed as

$$B_p \pm z\left(1-\frac{\alpha}{2}\right)\cdot \sigma\left(B_p\right), \tag{11}$$

where z is the inverse of the cdf that converts the probability value of $\left(1-\frac{\alpha}{2}\right)$ to a critical *z* cutoff score. For the 95% analytic-*z* CI, the lower and upper limits $\approx B_p \pm 1.96 \cdot \sigma\left(B_p\right)$. Further, researchers often use the inverse cumulative *t* distribution with degrees of freedom (*df*) equal to $n-2$ for estimating the *SE*. Hence, the $(1-\alpha)\cdot 100\%$ analytic-*t*, symmetrical CI surrounding β_p can be constructed as

$$B_p \pm t\left(1-\frac{\alpha}{2}\right)\cdot \sigma\left(B_p\right), \tag{12}$$

where t is the inverse of the cdf that converts the probability value of $\left(1-\frac{\alpha}{2}\right)$ to a critical *t* cut-off score based on $df = n-2$.

Simulation

Design

Distribution (∅)

Five distributions were evaluated (Figure 1). First, the X and Y scores follow BLNC that is assumed for the estimation of Pearson's correlation r in Equation 1. This distribution expects to produce an accurate r estimation, which can appropriately be converted to CL_r in Equation 1. The proposed B_p is also expected to be accurate. The remaining distributions include four types of common symmetrical distributions in behavioral research—PBS-normal distribution, t distribution (with $df = 18$), uniform distribution, and beta distribution (with alpha = beta = 0.5)—that adhere to the PBS function for generating the PBS-based X and Y scores. Figure 1 includes all these five different bivariate distributions, and rows 2 – 5 show that researchers may easily miss that the x and y are indeed (PBS-based) related if they obtain and evaluate r in their data-analytic plan.

Sample size (n)

Six levels of sample sizes—20, 50, 100, 300, 500, and 1000—were evaluated, which comprehensively cover a small to large sample size in behavioral research.

Population PBS (β_p)

Nine levels of β_p — .50, .55, .60, .65, .70, .75, .80, .85, and .90 — were examined. These values are comprehensive in covering most levels of ES in practice.

These factors are combined to produce a design with $5 \times 6 \times 9 = 270$ conditions. Each condition was replicated 1,000 times to evaluate the accuracy of B_p. For the bootstrap CIs, 2,000 samples were resampled with replacement to generate the BSI-z, BSI-t, BPI, and BCaI. The simulation was conducted in RStudio (2020), and the code is presented in Supplementary Materials below.

Data Generation Procedure

For the first type of distribution (BLNC), X scores were generated from a normal distribution, $N(0, 1^2)$. The linear-related Y scores were generated from

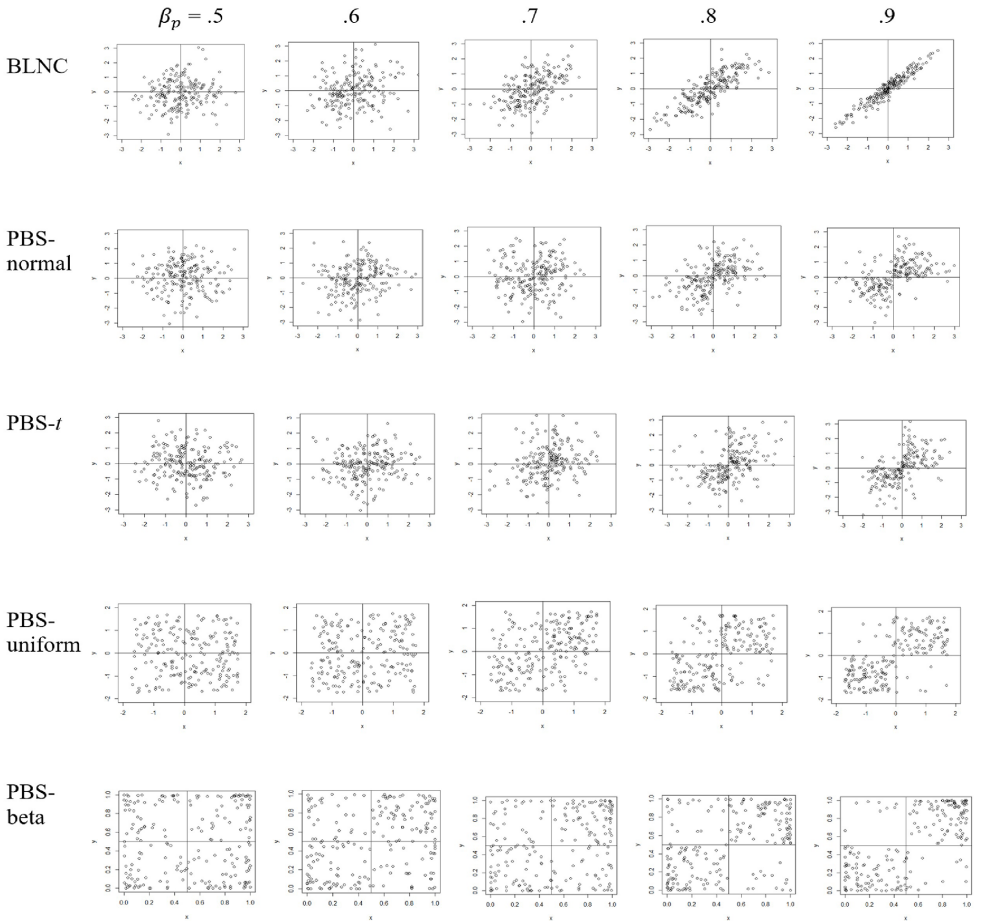
$$Y = \rho X + e_Y \tag{13}$$

where ρ is the population Pearson's correlation r converted from the population PBS, β_p through Equation 1, and e_Y is the error score generated from a $N(0, 1 - \rho^2)$. Given this method, X and Y are expected to be linearly correlated with a level of ρ.

For the remaining distributions, the simulation was executed in the R package (truncdist; Nadarajah, & Kotz, 2006) that can generate truncated data for most commonly found probability distributions. This means that when a generated X score is above (or

Figure 1

Scatterplots for 200 Simulated (x, y) Points That Come From BLNC, PBS-Normal, PBS-t, PBS-Uniform, and PBS-Beta Distributions



Note. β_p is the population probability-of-bivariate-superiority (PBS) value. BLNC refers to the bivariate linear, normal, and continuous distribution, PBS-normal is the PBS-based normal distribution, PBS-*t* is the PBS-based *t* distribution, PBS-uniform is the PBS-based uniform distribution, and PBS-beta is the PBS-based beta distribution.

below) the mean of all the other X scores, the package can generate a Y score that is above (or below) the true population mean of a probability distribution (i.e., PBS-normal distribution, t distribution, uniform distribution, and beta distribution). Specifically, a sample B_p was first generated from a binominal distribution, $B(n, \beta_p)$, to allow sampling

distributions of the PBS values. Second, the X scores were generated from one of the four symmetrical distributions: PBS-normality, $N(0, 1^2)$; t distribution, $t(18)$; uniform distribution, $U(-\sqrt{12}/2, \sqrt{12}/2)$; and beta distribution, Beta(0.5, 0.5). Third, when a generated X score was above (or below) the mean of all other X scores, there is a B_p likelihood (based on the binomially generated B_p) that a simulated Y score would be above (or below) the criterion or population mean (i.e., 0 for PBS-normality, 0 for t distribution, 0 for uniform distribution, and 0.5 for beta distribution, respectively) of a truncated probability distribution. Consequently, the generated Y scores followed normality, t distribution, uniform distribution, and beta distribution, and there was a B_p likelihood that when the X score was above (or below) the mean of all other X scores, the Y score would also be above (or below) the mean of all other Y scores. Once the data were simulated, the 6 CIs were constructed for comparisons.

Evaluation Criteria

Bias

Bias is used to evaluate the performance of the point estimates for the true PBS (β_p), i.e., bias $= \overline{B_p} - \beta_p$, where $\overline{B_p}$ is the mean of the 1,000 replicated B_p estimates, respectively.

Coverage Probability (CP)

Coverage probability is defined as the likelihood that the 95% CIs surrounding the B_p could span the true associated value (i.e., β_p) across 1,000 replications. That is,

$$\text{CP} = \sum_{i=1}^{1,000} \#\left[l(i) < \beta_p \wedge u(i) > \beta_p\right] / 1{,}000 \tag{14}$$

where $\#\left[l(i) < \beta_p \wedge u(i) > \beta_p\right]$ is the count function that count the number of times that the lower limit is smaller than β_p and the upper limit is larger than β_p. For the 95% CI, the expected CP should ideally be .95. To allow sampling error for CP, Chan and Chan (2004) suggested that a 95% CI should be regarded as acceptable, when the associated CP is within the range of [.925, .975].

Width of the CI

The width is defined as the difference between the upper and lower limits of a CI. A narrower (or wider) CI means that the method can produce a more (or less) precise boundary surrounding the B_p estimate, but this CI should also maintain a good CP in order to be regarded as an appropriate method. This is because an overly precise CI tends to decrease the likelihood that the CI could span the true parameter value, whereas an overly wide CI would, in theory, result in a CP close to 100%, but this could be too wide without any practical inferences.

Type 1 Error and Power

When $\beta_p = .50$ (i.e., lack of effect), Type 1 error is used to evaluate the chance that a constructed 95% CI does not span this true value across replicated samples, leading to an error in making a statistical inference. Of the 1,000 replications, the nominal number of the CIs that does not span the value of .50 should be as close as possible to 50 (or 5% of the 1,000 replicated samples). When $\beta_p > .50$, Power is used to evaluate the chance that a constructed 95% CI does not span the value of .50 across the replicated samples, which yields a significant result and correct decision.

Simulation Results

Bias

When the assumption of BLNC was met, B_p produced good results (see Table 1). The biases ranged from -.0069 to .0042 with a mean of .0001, meaning that B_p is highly accurate in quantifying the level of PBS for data that follows the conventional BLNC distribution. When data followed the PBS-based distributions, B_p performed equally well. For PBS-normal, the biases ranged from -.0428 to .0023 with a mean of -.0103 (range = [-.0428, .0023], mean = -.0103). For PBS-*t*, range = (-.0435, .0005), and mean = -.0112. For PBS-uniform, range = (-.0306, .0032), and mean = -.0080. For PBS-beta, range = (-.0243, .0012), and mean = -.0056.

Table 1

Biases of B_p When Data Followed 5 Types of Distributions: BLNC, PBS-Normal, PBS-t, PBS-Uniform, and PBS-Beta

Bias	BLNC	PBS-normal	PBS-*t*	PBS-uniform	PBS-beta	Overall
M	.0001	-.0103	-.0112	-.0080	-.0056	-.0070
SD	.0018	.0099	.0104	.0076	.0056	.0087
Min	-.0069	-.0428	-.0435	-.0306	-.0243	-.0435
Max	.0042	.0023	.0005	.0032	.0012	.0042

Note. BLNC = the bivariate linear, normal, and continuous distribution; PBS-normal = the PBS-based normal distribution; PBS-*t* = the PBS-based *t* distribution; PBS-uniform = the PBS-based uniform distribution; PBS-beta = the PBS-based beta distribution.

CP, Width, Type 1 Error and Power

Of the six methods, the BSI-*z* and BSI-*t* have the largest chance of spanning the true parameter value (see Table 2). The CPs ranged from .9440 to .9940 with a mean of .9668 for BSI-*z*, and they ranged from .9460 to .9980 with a mean of .9695 for BSI-*t*. On the other hand, a CI that produces an overly large CP (i.e., CP > .95) does not necessarily mean that this CI is the most accurate. The over-coverage of the true parameter value is in part

due to the unnecessarily wide and imprecise CI that can always span the true parameter value, which in turn leads to inaccurate Type 1 error and Power. In this case, the widths of the BSI-*z* ranged from .0399 to .5130 with a mean of .2029, and the widths of the BSI-*t* ranged from .0399 to .5499 with a mean of .2104, which are the widest relative to all the other methods. The means of the Type 1 error rates were .0380 and .0359 for BSI-*z* and BSI-*t*, respectively, which are smaller than the nominal value of .05, meaning that both methods are overly conservative in rejecting the null hypothesis. However, the means of the Power rates were .7895 and .7819 for BSI-*z* and BSI-*t*, respectively, and they were higher (or lower) than the means of the power rates produced by the BPI and BCaI (or analytic-*z* and analytic-*t*).

Table 2

Performance of the 6 Different CIs: Analytic-z, Analytic-t, BSI-z, BSI-t, BPI, and BCaI

Performance	Analytic-*z*	Analytic-*t*	BSI-*z*	BSI-*t*	BPI	BCaI
CP						
M	.9326	.9371	.9668	.9695	.9458	.9278
SD	.0219	.0195	.0100	.0115	.0410	.0218
Min	.8120	.8670	.9440	.9460	.7550	.8130
Max	.9640	.9700	.9940	.9980	1.0000	.9650
% within [.925, .975]	.7593	.8111	.7926	.6741	.6370	.6926
Width						
M	.1695	.1755	.2029	.2104	.2009	.1978
SD	.1151	.1241	.1465	.1581	.1433	.1436
Min	.0372	.0372	.0399	.0399	.0399	.0401
Max	.4274	.4581	.5130	.5499	.5064	.4986
Type 1 Error						
M	.0547	.0547	.0380	.0359	.0225	.0509
SD	.0101	.0101	.0075	.0093	.0127	.0072
Min	.0360	.0360	.0230	.0150	.0010	.0320
Max	.0740	.0740	.0500	.0490	.0460	.0660
Power						
M	.8108	.8108	.7895	.7819	.7378	.7361
SD	.2967	.2967	.3151	.3228	.3631	.3548
Min	.0580	.0580	.0380	.0320	.0040	.0280
Max	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Note. *M* = mean; *SD* = standard deviation; BSI-*z* = bootstrap standard interval based on the analytic-*z* *SE* approach: BSI-*t* = bootstrap standard interval based on the analytic-*t* *SE* approach; BPI = bootstrap percentile interval; BCaI = bootstrap bias-corrected and accelerated interval; % within [.925, .975] = the percentage of the estimated coverage probabilities that fell within [.925, .975] across the 270 simulation conditions; CP = coverage probability.

For the two percentile-based bootstrap CIs, BCaI has the best protected Type 1 error rates, which ranged from .0320 to .0660, with a mean of .0509: very close to the nominal value of .05. On the other hand, the Power of BCaI was found to be the smallest (mean = .7361) relative to all other methods. This finding is understandable because a better protected Type 1 error rate tends to decrease the Power in detecting a significant result. Of the 270 conditions, 187 (or 69.26%) of the CPs fell within the criterion of [.925, .975], but the mean of the CPs was the smallest (.9278) relative to all the other methods. Comparatively, another percentile-based CI (BPI) is more conservative than BCaI, when a researcher is making an inferential-statistical decision. Here, the mean Type 1 error rate was .0225, and the mean Power rate was .7378; both values were small relative to the other methods. Of the 270 conditions, 172 (or 63.70%) of the CPs fell within the criterion of [.925, .975], although the mean CP (.9458) was the closest to the nominal value of .95.

Regarding the two analytic approaches, both the analytic-*z* and analytic-*t* methods produced the largest means of the Power rates (.8108), and their means of the Type 1 error rates (.0547) were slightly larger than the nominal value of .05, meaning that both methods are slightly liberal in detecting a significant result. The widths of the CI were the narrowest (or the most precise) relative to the other methods. That is, the widths ranged from .0372 to .4274 with a mean of .1695 for the analytic-*z*. The use of the *t* distribution in constructing the analytic-*t* made the widths slightly wider than the analytic-*z*, and they ranged from .0372 to .4581 with a mean of .1755. The wider analytic-*t* approach improved the performance of the CPs. Of the 270 conditions, 219 (or 81.11%) produced a CP within the criterion of [.925, .975], which is the largest among all the other methods. For the analytic-*z* method, 205 (or 75.93%) conditions resulted in a CP that fell within the criterion of [.925, .975].

Effects of Sample Sizes on CP, Width, Type 1 Error, and Power

Given that sample size is the only factor that researchers can plan and control in practice, this section examines the effects of different sample size levels on the 6 CIs[1] (see Table 3).

1) There was no obvious difference regarding the effects of the data distributions and true β_p values on the CP and width, and hence, these effects are not explained. Moreover, data distribution did not influence the Type 1 error and Power rates of the CIs, and thus the explanation for these effects are excluded. On the other hand, different levels of sample sizes and β_p values were found to affect the performance of the 6 CIs, and hence, their influences are further discussed in the following sections.

Table 3

Effects of Different Sample Sizes on the Coverage Probability, Confidence Width, Type 1 Error, and Power of the 6 Different CIs

Performance / *n*	Analytic-*z*	Analytic-*t*	BSI-*z*	BSI-*t*	BPI	BCaI
CP						
20	.9180	.9336	.9744	.9835	.9700	.8899
50	.9314	.9381	.9724	.9755	.9634	.9261
100	.9364	.9386	.9675	.9699	.9466	.9267
300	.9374	.9381	.9650	.9656	.9374	.9396
500	.9364	.9380	.9621	.9625	.9318	.9417
1000	.9358	.9362	.9596	.9597	.9256	.9427
Width						
20	.3746	.4015	.4774	.5118	.4672	.4684
50	.2415	.2478	.2880	.2955	.2863	.2795
100	.1712	.1734	.1978	.2003	.1974	.1917
300	.0989	.0993	.1104	.1109	.1105	.1073
500	.0765	.0767	.0845	.0847	.0847	.0823
1000	.0541	.0541	.0590	.0591	.0592	.0576
Type 1 Error						
20	.0404	.0404	.0302	.0246	.0046	.0544
50	.0668	.0668	.0332	.0298	.0120	.0494
100	.0536	.0536	.0370	.0344	.0180	.0492
300	.0560	.0560	.0396	.0392	.0294	.0492
500	.0578	.0578	.0434	.0432	.0334	.0534
1000	.0534	.0534	.0446	.0444	.0374	.0500
Power						
20	.0678	.0678	.0500	.0370	.0092	.0350
50	.1282	.1282	.0764	.0680	.0320	.0548
100	.1784	.1784	.1372	.1320	.0866	.1104
300	.4146	.4146	.3696	.3674	.3194	.3280
500	.6000	.6000	.5604	.5592	.5174	.5252
1000	.8834	.8834	.8696	.8690	.8540	.8502

Note. *n* = sample size; BSI-*z* = bootstrap standard interval based on the analytic-*z* *SE* approach; BSI-*t* = bootstrap standard interval based on the analytic-*t* *SE* approach; BPI = bootstrap percentile interval; BCaI = bootstrap bias-corrected and accelerated interval; CP = coverage probability.

First, when sample size was increased, the CPs obtained from the 6 CIs tended to be closer to the nominal value of .95. Specifically, the analytic-*z*, analytic-*t*, and BCaI started from a smaller mean CP (.9180, .9336, and .8899, respectively) when $n = 20$, and it increased to a value closer to .95 (.9358, .9362, and .9427, respectively) when $n = 1{,}000$. Comparatively, the BSI-*z*, BSI-*t*, and BPI began with a larger mean CP

(i.e., .9744, .9835, and .9700, respectively) when n = 20, and it decreased to a value closer to .95 (i.e., .9596, .9597, and .9256, respectively) when n = 1,000. It is noteworthy that the mean CP yielded by the BPI always decreased when n increased.

Second, the mean widths of the 6 CIs became narrower when n increased. The differences of the mean widths of the 6 CIs were the most obvious when n = 20, with the narrowest mean width of .3746 for the analytic-z, and the widest mean width of .5118 for the BSI-*t*. However, the mean widths yielded by all the 6 CIs, which ranged from .0541 to .0592, became highly similar when n = 1,000.

Third, BCaI always led to the best protected mean Type 1 error rates (ranging from .0492 to .0544), and the analytic-*z* and analytic-*t* produced reasonable mean error rates (both ranged from .0404 to .0668). All these 3 methods tended to produce a mean error rate close to .05 when n increased. The remaining methods (BSI-*z*, BSI-*t*, and BPI) had a conservative mean Type 1 error rate (.0302, .0246, and .0046) when n = 20, and it became slightly closer to .05 (.0446, .0444, and .0374) when n = 1,000.

Fourth, when n increased, the mean Power rates also increased for the 6 CIs. When n was small, there were noticeable differences in the mean Power rates (e.g., .1282 for the analytic-*z* or analytic-*t* and .0320 for the BPI when n = 50), but all the mean Power rates were larger than 85% when n = 1,000.

Effects of the β_p Values on Power

The only influential factor remaining lies in the effect of different β_p values on the Power rates of the 6 Cis (see Table 4). As expected, when β_p increased from .55 to .90, the mean Power rates of the 6 CIs increased accordingly. The most powerful methods were the analytic-*z* and analytic-*t*, and they shared the same Power rates, which increased from .3787 (when β_p = .55) to .9917 (when β_p = .90). The percentile-based BPI and BCa were relatively the least powerful. For BPI, the Power rates ranged from .3013 to .9460 with a mean of .7378. For BCaI, the Power rates ranged from .3173 to .9377 with a mean of .7361. For the remaining bootstrap-analytic approaches, the range of the Power rates was [.3439, .9863] with a mean of .7895 for BSI-z, whereas the range was [.3388, .9825] with a mean of .7819 for BSI-t. Generally, increasing the β_p values tends to produce a similar and comparable increase in the Power rate for all the 6 methods, and their slight differences depend upon whether they have a more conservative (or liberal) Type 1 error rate at the baseline when β_p = .50.

In sum, when a study sample is small (n = 20), the analytic-*t* appears to be the most appropriate CI with a good CP, large Power, reasonable Type 1 error, and generally narrow and precise width of the CI. When a study sample is large (n = 1,000), a scenario in which many inferential statistics may be overly sensitive and powerful in signalling a significant result leading to an inflated Type 1 error, the BCa is the most desirable choice because of its highly protected mean Type 1 error and reasonable Power rates. It also has

Table 4

Effects of the β_p Levels on the Power of the 6 CIs

β_p	Analytic-*z*	Analytic-*t*	BSI-*z*	BSI-*t*	BPI	BCaI
.55	.3787	.3787	.3439	.3388	.3031	.3173
.60	.6402	.6402	.6047	.5975	.5580	.5643
.65	.7740	.7740	.7402	.7302	.6811	.6833
.70	.8601	.8601	.8340	.8233	.7718	.7701
.75	.9146	.9146	.8986	.8885	.8390	.8310
.80	.9509	.9509	.9407	.9328	.8849	.8774
.85	.9761	.9761	.9678	.9612	.9183	.9074
.90	.9917	.9917	.9863	.9825	.9460	.9377

Note. β_p = the true population probability-of-bivariate-superiority (PBS) value; BSI-*z* = bootstrap standard interval based on the analytic-*z SE* approach; BSI-*t* = bootstrap standard interval based on the analytic-*t SE* approach; BPI = bootstrap percentile interval; BCaI = bootstrap bias-corrected and accelerated interval.

a good mean CP and moderate width of CI for maintaining a good balance between Type 1 error and Power.

Real-World Example

On the basis of sensation-seeking theories, Kalichman and Rompa (1995) developed a scale, called the Sexual Compulsivity Scale (SCS), which measures whether people possess higher levels of sexual compulsivity or oriented thinking in their daily lives. The SCS is a 10-item inventory and participants respond on a 4-point Likert scale (from 1 "*not at all like me*" to 4 "*very much like me*"). Sample questions include "I find myself thinking about sex while at work," and "my desire to have sex has disrupted my daily life." The SCS has been used and validated in many studies (e.g., Gaither, Sellbom, & Meier, 2003; Humphreys & Brousseau, 2010; Milhausen, Graham, Sanders, Yarber, & Maitland, 2010).

There is an open-access database that provides a raw SCS data-set for research purposes (the data used for the current analysis is available in the Supplementary Materials). This database saved n = 3,375 valid respondents (with 1 missing value), who provided their self-report scores on SCS. A common research question associated with this data set involves whether or not respondent's age is related to sexual compulsivity. Researchers typically compute a total score of the 10 items to reflect the level of sexual compulsivity, and this variable is approximated as a continuous variable with a score ranging from 4 to 40. Age is measured in terms of years, which is also a continuous variable. The conventional Pearson's correlation showed that r = .0037, 95% CI [-.0300, .0374], meaning that only .00137% (r^2 = .0000137) of variance of sexual compulsivity can be accounted for by

age. The 95% CI also spans the value of 0, meaning that the correlation is not significant at the .05 level. Comparatively, if a researcher examines the relationship based on PBS (see dataset in Supplementary Materials), then the result will be interpreted differently with B_p = .5170, and the 95% BCaI [.5001, .5342] (or the 95% analytic-*t* CI [.5001, .5339]). Note that although the B_p value is not large (.5170), this effect is significant at the .05 level, and the 95% CI does not span the value of .50. One can thus interpret the result as being a 51.7% likelihood that when a person is older than the mean age of all other participants (31.02 years), the person will also have a sexual compulsivity higher than the mean sexual compulsivity score (23.45 in SCS) of all other participants.

Conclusion and Discussion

A lack of bivariate linear correlation does not imply a lack of bivariate relationship. Most behavioral researchers examine a research hypothesis that is based on linear relationships between variables. They typically specify and choose a linear-based statistical model (e.g., Pearson's correlation), despite the fact that there are many other types of bivariate relationships (e.g., curvilinearity; Li, 2018b). Dunlap (1994) is arguably one of the pioneer studies which attempted to develop a method to evaluate the level of PBS instead of bivariate linearity. Yet, Dunlap's approach neither extends the concept of bivariate relationship beyond linearity nor provides a *SE* and CI algorithm for estimating the associated sampling error and precision. Li and Waisman (2019) mathematically derived an algorithm for estimating PBS (B_p) between two continuous variables, but did not provide the mathematical details for developing an analytic method for the *SE* and CI. Given that many behavioral researchers would prefer a PBS-based interpretation (Brooks et al., 2014), and PBS-based relationships have tremendous potential for explaining many bivariate relationships that may have been missed in previous research, this study is an important piece of work that fills in this research gap.

The present study is a crucial development in extending and promoting the use of PBS in practice. Researchers are increasingly aware of the importance and usefulness of PBS in examining relationships between variables. Conceptually, both *r* and PBS can be used to measure and quantify the level of bivariate relationships that may exist in two variables. Pearson's correlation is arguably the most widely employed statistical measure because of its simple, easy-to-understand concept that linearity is the underlying explanation for why two variables covary together. In addition to *r*, there are indeed many other alternative methods for detecting nonlinear bivariate relationships, but these methods have their own limitations. For instance, Reshef et al. (2011) proposed a new correlational estimate (maximal information coefficient [MIC]) that can potentially detect 27 different types of bivariate relationships (e.g., linear, cubic, parabolic). One potential weakness to this approach is that MIC is too generic in identifying any particular type of relationship, and researchers may find it difficult to interpret the value of MIC because, for example,

a value of MIC = 0.5 could refer to very different meanings or sizes of an effect for any of the 27 types of relationships. Moreover, as aptly noted by a reviewer of this study, the Power of the MIC in detecting any particular type of relationship may suffer. PBS, on the other hand, was proposed to answer a slightly different question that many psychological researchers address. That is, researchers should test models for detecting ordinal relationships because many variables used in psychological research are indeed ordinal-scale (e.g., 5-point Likert scale). When researchers analyze ordinal data with (inappropriate) metric models, Liddell and Kruschke (2018) found that there will be an increase in Type 1 error and a loss of Power. PBS is a measure that is not only easier to be interpreted than *r*, but it also assesses an ordinal relation between *x* and *y* without depending upon the unnecessary linearity assumption.

Most publication manuals in psychology (e.g., American Psychological Association, 2010) require that researchers report the CI and interpret the significance level in addition to a point estimate of a statistical measure. This study extends early studies (e.g., Blomqvist, 1950) to develop a mathematical foundation for the PBS estimate (B_p) and an analytic method for estimating the associated *SE*, and comparing both the analytic and bootstrap approaches to the CI constructions for B_p. The present study provides an important mathematical proof for estimating the sampling error and precision for B_p, so that this area of research can further be tested and used by theoretical and applied researchers. Moreover, this proof also clearly shows that PBS can be conceptualized as an independent statistical model, which does not need to be converted from *r* based on the assumption of BLNC in Dunlap (1994).

Our simulation results show that each of the 6 CI methods behave in a manner that may serve for different research purposes. If a researcher prefers to use a method that encompasses many potential study effects, perhaps in the early stages of exploratory research, then the researcher can use the analytic-*t* CI because it has the highest Power in signalling a significant B_p result with a slightly liberal Type 1 error rate (i.e., mean Type 1 error rate = .0547 in the current simulation). On the other hand, when a researcher is interested in confirming a study effect that was found and published by other researchers in earlier studies, the BCaI method is the most desirable because it has the most accurate Type 1 error rate (i.e., mean Type 1 error rate = .0509 in the current simulation), a criterion that should be stringently protected in the later stages of a research study. Another criterion for choosing between the analytic and bootstrap CIs is the sample size in a study. As shown in the simulation, the BCaI tends to behave well in terms of Type 1 error, Power, CP, and width with a large sample (e.g., n = 1,000), whereas the analytic-*t* CI possesses good CP, Power, precise confidence width, and reasonable (and slightly liberal) Type 1 error with a small sample (e.g., n = 20).

In the real-world example, the results led to different conceptual understandings of the pattern of relationships based on *r* and B_p. The correlation between age and sexual compulsivity was .0037 (which is close to 0), implying that linearity should not

be the underlying pattern or shape that governs the association between age and sexual compulsivity. On the other hand, if one uses B_p to conceptualize their association, there is a 51.7% likelihood that a person who is older than the mean age of all other participants (31.02 years), will also have a sexual compulsivity higher than the mean sexual compulsivity level (23.45 in SCS) of all other participants. Furthermore, one would have a different statistical inference if one uses the CI for r (non-significant result) and the CI for B_p (significant result).

In terms of theory, this study provides the details of the necessary mathematical proof (Equations 3 - 12) for the point estimate, *SE* and CI for B_p, as well as the code for future researchers who are interested in investigating PBS-based relationships. One future direction involves generalizing PBS to complex relationships (e.g., the conditional effect of X on Y controlling for covariates, interaction, mediation, and moderation) that are frequently found in behavioral research. The current mathematical proof lays a foundation for theoretical researchers to extend and develop PBS. Another direction is examining additional factors that could influence the behavior of B_p as well as its *SE* and CI. As in other simulations, the present study cannot include all different factors and examine their impact on a statistical method. For example, the numbers of $a(x_i, y_i)$ points and $c(x_i, y_i)$ points should be similar (i.e., symmetric distribution) as assumed in the classical theory of likelihood-based relationships in Blomqvist (1950). Future research could examine whether B_p is robust to asymmetric distributions (e.g., $ln(\mu, \sigma^2)$) for X and Y, with the expected population means of X and Y (e.g., $e^{(\mu + \sigma^2/2)}$) serving as the corresponding cut-off criteria that govern PBS between X and Y.

Funding: This research was funded by a University Research Grants Program (URGP) to Johnson Ching-Hong Li in the Department of Psychology at the University of Manitoba (#47094).

Acknowledgments: The authors have no support to report.

Competing Interests: The authors have declared that no competing interests exist.

Data Availability: Data for this article is freely available (see Kalichman & Rompa, 1995).

Supplementary Materials

For this article the following Supplementary Materials are available (for access see Index of Supplementary Materials below):

- Via the PsychArchives repository: Simulation code.
- Via the OpenPsychometrics repository: Raw dataset.

Index of Supplementary Materials

Li, J. C.-H., & Tze, V. M. C. (2021). *Supplementary materials to: Analytic and bootstrap confidence intervals for the common-language effect size estimate* [Code]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.4720

Kalichman, S. C., & Rompa, D. (1995). *Answers to the Sexual Compulsivity Scale from Kalichman and Rompa* [Dataset]. OpenPsychometrics. https://openpsychometrics.org/_rawdata/SCS.zip

References

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

Blomqvist, N. (1950). On a measure of dependence between two random variables. *Annals of Mathematical Statistics, 21*(4), 593-600. https://doi.org/10.1214/aoms/1177729754

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *The Journal of Applied Psychology, 99*(2), 332-340. https://doi.org/10.1037/a0034745

Chan, W., & Chan, W.-L. (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. *Psychological Methods, 9*(3), 369-385. https://doi.org/10.1037/1082-989X.9.3.369

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*(3), 494-509. https://doi.org/10.1037/0033-2909.114.3.494

Cliff, N. (1994). Predicting ordinal relations. *British Journal of Mathematical & Statistical Psychology, 47*(1), 127-150. https://doi.org/10.1111/j.2044-8317.1994.tb01028.x

Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research, 31*(3), 331-350. https://doi.org/10.1207/s15327906mbr3103_4

Cliff, N., & Keats, J. (2003). *Ordinal measurement in the behavioral sciences*. Mahwah, NJ, USA: Lawrence Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ, USA: Erlbaum

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29. https://doi.org/10.1177/0956797613504966

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods, 11*(3), 217-227. https://doi.org/10.1037/1082-989X.11.3.217

Dunlap, W. P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin, 116*(3), 509-511. https://doi.org/10.1037/0033-2909.116.3.509

Gaither, G. A., Sellbom, M., & Meier, B. P. (2003). The effect of stimulus content on volunteering for sexual interest research among college students. *Journal of Sex Research, 40*(3), 240-248. https://doi.org/10.1080/00224490309552188

Humphreys, T. P., & Brousseau, M. M. (2010). The sexual consent scale—revised: Development, reliability, and preliminary validity. *Journal of Sex Research, 47*(5), 420-428. https://doi.org/10.1080/00224490903151358

Kalichman, S. C., & Rompa, D. (1995). Sexual sensation seeking and sexual compulsivity scales: Reliability, validity, and predicting HIV risk behavior. *Journal of Personality Assessment, 65*(3), 586-601. https://doi.org/10.1207/s15327752jpa6503_16

Li, J. C.-H. (2016). Effect size measures in a two independent-samples case with non-normal and non-homogeneous data. *Behavior Research Methods, 48*, 1560-1574. https://doi.org/10.3758/s13428-015-0667-z

Li, J. C.-H. (2018a). Probability-of-superiority SEM (PS-SEM)—Detecting probability-based multivariate relationships in behavioral research. *Frontiers in Psychology, 9*, Article 883. https://doi.org/10.3389/fpsyg.2018.00883

Li, J. C.-H. (2018b). Curvilinear moderation—A more complete examination of moderation effects in behavioral sciences. *Frontiers in Applied Mathematics and Statistics, 4*, Article 7. https://doi.org/10.3389/fams.2018.00007

Li, J. C.-H., & Waisman, R. (2019). Probability of bivariate superiority: A non-parametric Common-Language statistic for detecting bivariate relationships. *Behavior Research Methods, 51*, 258-279. https://doi.org/10.3758/s13428-018-1089-5

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328-348. https://doi.org/10.1016/j.jesp.2018.08.009

May, H. (2004). Making statistics more meaningful for policy research and program evaluation. *The American Journal of Evaluation, 25*(4), 525-540. https://doi.org/10.1177/109821400402500408

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*(2), 361-365. https://doi.org/10.1037/0033-2909.111.2.361

Milhausen, R. R., Graham, C. A., Sanders, S. A., Yarber, W. L., & Maitland, S. B. (2010). Validation of the sexual excitation/sexual inhibition inventory for women and men. *Archives of Sexual Behavior, 39*, 1091-1104. https://doi.org/10.1007/s10508-009-9554-y

Nadarajah, S. & Kotz, S. (2006). R programs for computing truncated distributions. *Journal of Statistical Software, 16*(2016). https://doi.org/10.18637/jss.v016.c02

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., . . . Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science, 334*(6062), 1518-1524. https://doi.org/10.1126/science.1205438

RStudio. (2020). RStudio: Integrated Development for R. RStudio [Computer software]. rStudio. Retrieved from http://www.rstudio.com/

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*(1), 19-30. https://doi.org/10.1037/1082-989X.13.1.19

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.