Behavior Research Methods (2019) 51:258–279
https://doi.org/10.3758/s13428-018-1089-5

Probability of bivariate superiority: A non-parametric common-language statistic for detecting bivariate relationships

Johnson Ching-Hong Li[1] · Rory M. Waisman[2]

Published online: 10 August 2018

Abstract

Researchers often focus on bivariate normal correlation (r) to evaluate bivariate relationships. However, these techniques assume linearity and depend on parametric assumptions. We propose a new nonparametric statistical model that can be more intuitively understood than the conventional r: *probability of bivariate superiority* (PBS). Our development of B_p, the estimator of a PBS relationship, extends Dunlap's (1994) common-language transformation of r (CL_r) by providing a method to directly estimate PBS—the probability that when x is above (or below) the mean of all X, its paired y score will also be above (or below) the mean of all Y. Probability of superiority is an important form of bivariate relationship that until now could only be accurately estimated when data met the parametric assumptions for r. We specify the copula that forms the theoretical basis for PBS, provide an algorithm for estimating PBS from a sample, and describe the results of a Monte Carlo experiment that evaluated our algorithm across 448 data conditions. The PBS estimate, B_p, is robust to violations of parametric assumptions and offers a useful method for evaluating the significance of probability-of-superiority relationships in bivariate data. It is critical to note that B_p estimates a different form of bivariate relationship than does r. Our working examples show that a PBS effect can be significant in the absence of a significant correlation, and vice versa. In addition to utilizing the PBS model in future research, we suggest that this new statistical procedure be used to find theoretically important but previously overlooked effects from past studies.

Keywords Bivariate relationships · Correlation · Probability of superiority · Common language · Effect size

The need for statistical literacy—described as "the ability to interpret, critically evaluate, and communicate statistical information and messages" (Gal, 2002, p. 1)—is receiving growing worldwide attention. The United Nations Economic Commission for Europe (2009) published *A Guide to Improving Statistical Literacy*, a manual aimed at promoting statistical literacy among researchers, educators, businesses, policy makers, and the general public, and in 2010 the Royal Statistical Society launched *getstats*, a 10-year statistical literacy campaign.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-018-1089-5) contains supplementary material, which is available to authorized users.

✉ Johnson Ching-Hong Li
Johnson.Li@umanitoba.ca

[1] Lab for Research in Quantitative and Applied Statistical Psychology (LIQAS), Department of Psychology, University of Manitoba, P517B, Duff Roblin Building, Winnipeg, MB R3T 2N2, Canada

[2] School of Business, University of Alberta, Edmonton, Alberta, Canada

Despite calls for statistical literacy, most statistical models used today (e.g., correlation and its generalized forms, such as regression) in health, business, education, and behavioral research are difficult for many people to understand (e.g., Brooks, Dalal, & Nolan, 2014; Dunlap, 1994). Lack of statistical literacy is partly responsible. But the complexity of statistical models and the statistical jargon used by researchers must also bear some of the blame.

Correlational analysis is one of the most commonly used statistical techniques, and it exemplifies the latter problem. Although it is arguably among the least complex statistical models used by researchers, correlation is not conducive to intuitive understanding. Consider, for example, a correlational analysis that shows a significant positive correlation between physical exercise and mortality, with a correlation estimate of $r = .60$. The squared value, $r^2 = .36$, can be interpreted to mean that 36% of the variance in mortality can be explained by variance in physical activity. This description does not facilitate an intuitive mental picture of the important relationship between physical activity and mortality. The underlying statistical jargon, "proportion of variance explained," is not easily understood by

members of the public, policy makers, and public officials (May, 2004). Even some researchers in psychology may not be comfortable with this kind of statistical terminology (Brooks et al., 2014).

In addition, correlational analysis is problematic under many common conditions. Although the most commonly used method for estimating correlation, Pearson's *r*, is well-defined when the means and variances of *X* and *Y* are well-defined (e.g., Hogg & Craig, 1971), drawing a valid inference from the interpretation of *r* becomes less probable when the conditions of linearity, bivariate normality, and no outliers are violated (Onwuegbuzie & Daniel, 2002). Such condition violations are common in many research domains (e.g., engineering psychology: Bradley, 1982; educational and clinical psychology: Micceri, 1989), making correlational analysis, and parametric statistics in general, insufficient in many cases. Consequently, there have been increasing calls for the use of nonparametric statistics (Leech & Onwuegbuzie, 2002), going at least as far back as Siegal's (1956, p. vii) observation that "non parametric techniques of hypothesis testing are uniquely suited to the data of the behavioral sciences."

Researchers recognize multiple ways in which two continuous variables can be meaningfully related. Two general forms of relation are commonly evaluated: linearity and monotonicity. By definition, a *monotonic* relationship between *X* and *Y* indicates that the direction of change in *Y* when *X* changes is preserved across levels of *X*. Linear functions are monotonic. Other examples of monotonic functions are exponential function when the exponent is odd [e.g., $X = Y^3$, $X = Y^5$, or $X = \ln(Y)$]. It is generally suggested (e.g., Howell, 2013; Wilcox, 2012) that researchers use *r* for detecting linear relationships between *X* and *Y* and use nonparametric correlations (e.g., Spearman's rank correlation, r_s; Kendall & Gibbons, 1990) for detecting nonlinear monotonic relationships.[1]

Perhaps less well-known are the nonparametric methods available to estimate other forms of bivariate relationships. Generally, methods for detecting nonlinear monotonic bivariate relationships are even more complex and less understandable than correlation. Researchers need statistical methods that will allow them to make valid inferences from their data while still enabling them to communicate their analysis in a manner conducive to knowledge mobilization. What, then, is the solution?

We aim to contribute to such a solution by introducing a new statistic that can be more easily understood than correlation and is robust to common violations of parametric assumptions. *Probability of bivariate superiority* (PBS) is a nonparametric procedure for directly estimating the probability of superiority in a bivariate relationship between two continuous variables. The resulting effect size, B_p, is a common-language effect size estimate of the probability that a respondent who scores high (or low) on *X* will also score high (or low) on *Y*. Common-language effect sizes like B_p are interpreted using language that is more familiar to and more intuitively understood by people without a statistical background (Brooks et al., 2014).

Until now, nonparametric common-language effect sizes (Vargha & Delaney, 2000) have been applied mostly to between-group comparisons. McGraw and Wong (1992) introduced a common-language effect size (*CL*), interpreted as the probability that a score sampled at random from one group will be larger than a score sampled at random from the other group. Grissom (1994) coined the term *probability of superiority* to appropriately describe the relationship estimated by *CL*, and Vargha and Delaney introduced the nonparametric estimator *A* of this probability. Dunlap (1994) derived the formula to convert *r* to a common-language effect size (CL_r) that describes the probability that when an *x* score is above (or below) the mean of all *X*, its paired *y* score is above (or below) the mean of all *Y*. With these developments, the common-language probability-of-superiority approach has started to gain traction in psychology and other disciplines (e.g., biology: Ling & Nelson, 2014; education: Huberty, & Lowman, 2000).

In this article we introduce B_p, a nonparametric extension of CL_r. It is important to note that B_p depends upon neither linearity, as *r* does, nor monotonicity, as rank correlation does. Probability of superiority in a bivariate relationship can exist and can be appropriately interpreted independent of there being a linear or monotonic relationship between *X* and *Y*. B_p does not rely on the parametric assumptions upon which *r* and CL_r depend and that are commonly violated in real-world research. But like CL_r, it is an effect size that can be interpreted as a likelihood, making it easier to understand (Brooks et al., 2014).

Like correlation, PBS indicates correspondence between *X* and *Y* scores but does not imply causation. PBS does, however, make a practically important relationship between *X* and *Y* easier to interpret. For example, rather than trying to comprehend that 36% of the variance in mortality can be explained by variance in physical activity, it is easier to understand that there is a 70% likelihood that seniors who exercise more than 1 hour per day (national average) will live longer than 81 years (national average). Not only is this likelihood description easier for most people to understand, it also communicates concrete guidelines for a recommended course of action (exercise more than 1 hour daily) and why to do it (increased chance of living longer than 81 years). Practically speaking, understanding the nature of the relationship between lifestyle choices and health may increase seniors' willingness to make beneficial lifestyle changes.

The probability of superiority in bivariate relationships has received little attention, despite calls from researchers in other

[1] Technically speaking, even if $X = Y^3$, this does not indicate the absolute lack of a linear relationship between *X* and *Y*. One could describe the relationship as being approximately linear, although this would produce more error than using a cubic relationship or a robust correlation (e.g., rank correlation).

disciplines (e.g., Nelson, 2006). In the remainder of this article, we begin to fill this gap. In the sections that follow, we (a) discuss how linearity can be estimated by *r*, monotonicity estimated by nonparametric correlations, and PBS estimated by common-language effect sizes (CLESs); (b) review the common-language effect sizes and Dunlap's (1994) CL_r; (c) explain the copula for the probability of bivariate superiority (γ) that is estimated by CL_r when the assumptions for *r* are met; (d) introduce our common-language effect size, B_p, for estimating a PBS relationship; (e) describe the PBS algorithm we have developed to directly compute B_p as a robust estimator of γ; (f) describe our Monte Carlo experiment and the resulting evidence of the robustness of B_p; and (g) demonstrate the application of PBS using both a simulated dataset and real data. Finally, we offer a general discussion of the implications of PBS for past and future research in the behavioral and social sciences.

Linearity measured by Pearson's correlation coefficient *r*

In 1895, Karl Pearson (1895) presented the algorithm for a ground-breaking statistical concept, Pearson's correlation coefficient *r*, which can be used to assess the degree and direction of linear association between two continuous variables. Since its development, *r* and its generalized forms (e.g., multiple correlation, *R* in regression) have been widely employed in behavioral and social sciences. According to Rodgers and Nicewander (1988), *r* (and its related concepts) is undoubtedly one of the most revolutionary mathematical and statistical procedures in the 20th century. Many commonly used statistical models have been developed on the basis of the concept of *r*, including its generalized forms (e.g., multiple regression, structural equation modeling), robust forms (e.g., Spearman's correlation, Kendall's tau correlation), and its extended applications (e.g., mediation models, reliability assessment). In equation, *r* is an estimator for the linear association between *X* and *Y*, when *X* and *Y* are linearly related (Hogg & Craig, 1971), which can be expressed as

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \tag{1}$$

where ρ is the population correlation coefficient, $E[(X-\mu_X)(Y-\mu_Y)]$ is the expected value of the multiplicative scores of $(X-\mu_X)$ and $(Y-\mu_Y)$, μ_X is the population mean of *X*, μ_Y is the population mean of *Y*, σ_X is the population *SD* of *X*, and σ_Y is the population SD of *Y*. In practice, the population values are replaced with the sample values in Eq. (1) for estimating the sample correlation *r*. Possible values of *r* range from – 1 to + 1 (i.e., perfect-negative to perfect-positive linear correlation).

Linearity is the central property of the relationship that *r* describes. Theoretically speaking, and as is implied in Eq. 1, the correlation coefficient *r* is well-defined as long as the means and variances of *X* and *Y* are well-defined, regardless of their distributional characteristics. However, Hogg and Craig (1971) stated that the correlation coefficient *r* proves to be a useful estimator for linear relationship only for certain kinds of distributions of two random variables. Figure 1 shows scatterplots of 100 simulated *X* scores coming from normal, uniform, positively skewed, and negatively skewed distributions, and of the 100 simulated *Y* scores conditional on the *X* scores, based on the linear correlation value *r* = .80. Although the correlation coefficient *r* can reliably describe and measure how the *X* and *Y* scores concentrate in a line, in some scenarios (e.g., positively skewed *X*) the locations of extreme *X*–*Y* points on the *X*–*Y* plane may not be meaningful. This is because the linear relationship may be overstated, given that the majority of other points are relatively spread apart or unrelated, and there is a line that connects these points with a few outlier points on the *X*–*Y* plane. Hence, the correlation coefficient *r* is more useful when *X* and *Y* are symmetrically distributed (e.g., normal, uniform) than when *X* and *Y* are asymmetrically distributed, despite the fact that "the formal definition of ρ does not reveal this fact" (Hogg & Craig, 1971, p. 74).

Furthermore, when the relationship between *X* and *Y* is not linear (e.g., quadratic, cubic, quartic, or quintic), correlational analysis can lead researchers to inaccurate inferences. Figure 2 shows 10,000 normal *X* and normal *Y* that perfectly follow quadratic, cubic, quartic, and quintic bivariate relationships, respectively. For the quadratic data, *r* = .009; for cubic, *r* = .780; for quartic, *r* = .027; and for quintic, *r* = .512. If *r* is found to be statistically significant, a researcher may incorrectly infer that a linear relationship does exist (see note 1 above). If *r* is found to be not statistically significant, a researcher may incorrectly infer that there is no important relationship between the variables. This could explain why many applied statisticians and methodologists suggest that *r* should only be used to detect the direction (positive/negative) and magnitude of a linear relationship between *X* and *Y* when *X* and *Y* form a bivariate normal distribution (e.g., Howell, 2013).

Limited interpretability It took the publication of Cohen's (1988) highly influential text for meaningful interpretation of *r* to be appreciated in the dissemination of research findings. Cohen's guidelines used the concepts of proportion of variance explained and coefficient of determination, and he suggested rough categorical guidelines for the treatment of *r* as an effect size. For example, consider the relationship between academic achievement, measured using college GPA, and students' motivation. A computed correlation between these two variables, *r* = .30, can be interpreted by first computing its squared value (r^2 = .09). The r^2 value can in turn be interpreted

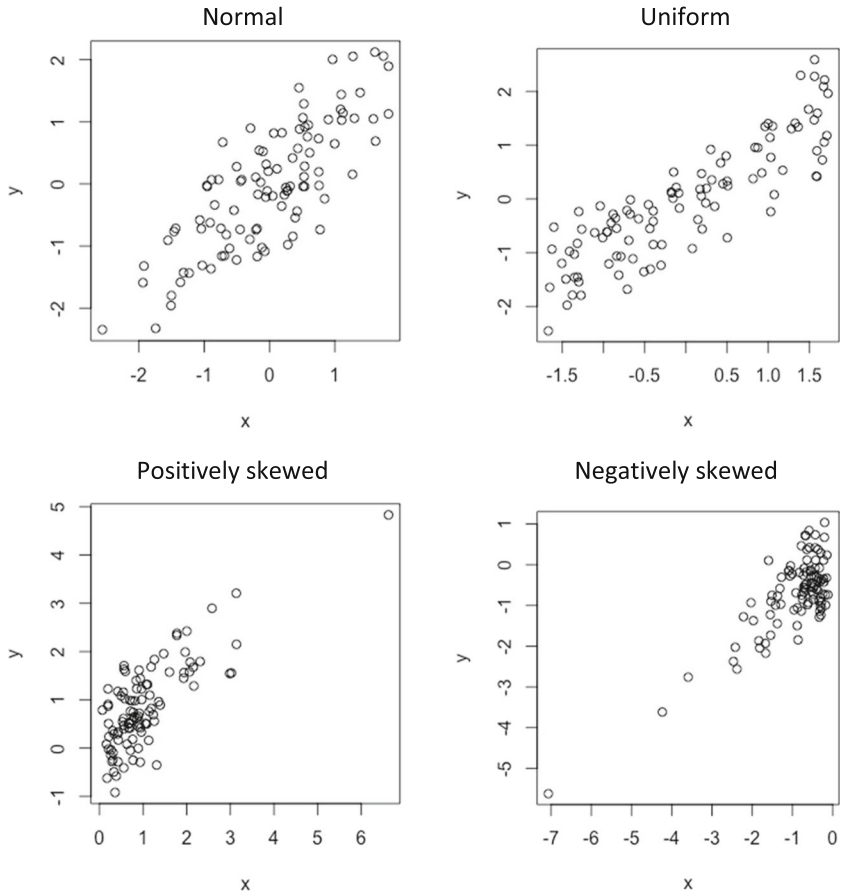

Fig. 1 Scatterplots for linear-based *X*–*Y* space with a normal, a uniform, a positively skewed, and a negatively skewed distribution, when the true correlation coefficient is .80

to mean that 9% of the variance in GPA can be explained by variance in motivation. In this case, .09 is the coefficient of determination. Despite Cohen's efforts, r^2 (or r) arguably remains one of the most confusing statistical concepts in behavioral and social research.

According to May (2004), three guidelines are essential for better disseminating statistical information: understandability, interpretability, and comparability. *Understandability* is enhanced when statistics are presented in plain language, without statistical jargon and assumptions. *Interpretability* requires the metric of a statistic to be familiar and easily understandable by the public. *Comparability* demands that a statistic be compared directly, without any need for manipulation and modification. Correlation meets this last requirement only, but it is certainly deficient in terms of understandability and interpretability.

In light of the difficulty with understanding and interpreting *r*, Brooks et al. (2014) conducted two experiments in which they recruited undergraduate students and asked them to rate statistical information on the basis of the three criteria of understandability, usefulness, and effectiveness. The statistical information was presented as (a) proportions of variance explained (or the coefficient of determination; r^2), (b) probability-based common-language effect sizes (*CL*), and (c) tabular binomial effect size displays (BESD). Participants perceived both *CL* and BESD as significantly more understandable and useful than r^2. Referring back to May's (2004) guidelines, it easy to appreciate why the proportion of variance explained was not preferred: It is difficult to understand because it is pure statistical jargon that is challenging to interpret.

The interpretative challenge is especially problematic when Cohen's (1988) effect size guidelines are applied. For example, $r^2 = .09$, interpreted as 9% of variance explained, is considered a medium effect size. But a person not fully comfortable with statistical terminology may justifiably conceive of 9% as a very low proportion, making any reference to it as a medium effect size confusing and perhaps even perceived to be misleading. Using a metric with these weaknesses can compromise attempts to disseminate findings. Despite the aforementioned weaknesses, *r* remains a commonly used measure of correspondence between *X* and *Y* scores.

When interpreted strictly as an indicator of directionality of *Y* to *X* correspondence, *r* is less difficult to understand. It is apparent from the sign of *r* whether *Y* increases or decreases as *X* increases. Unfortunately, such simplified interpretation has distinct disadvantages. First, it ignores the magnitude of the

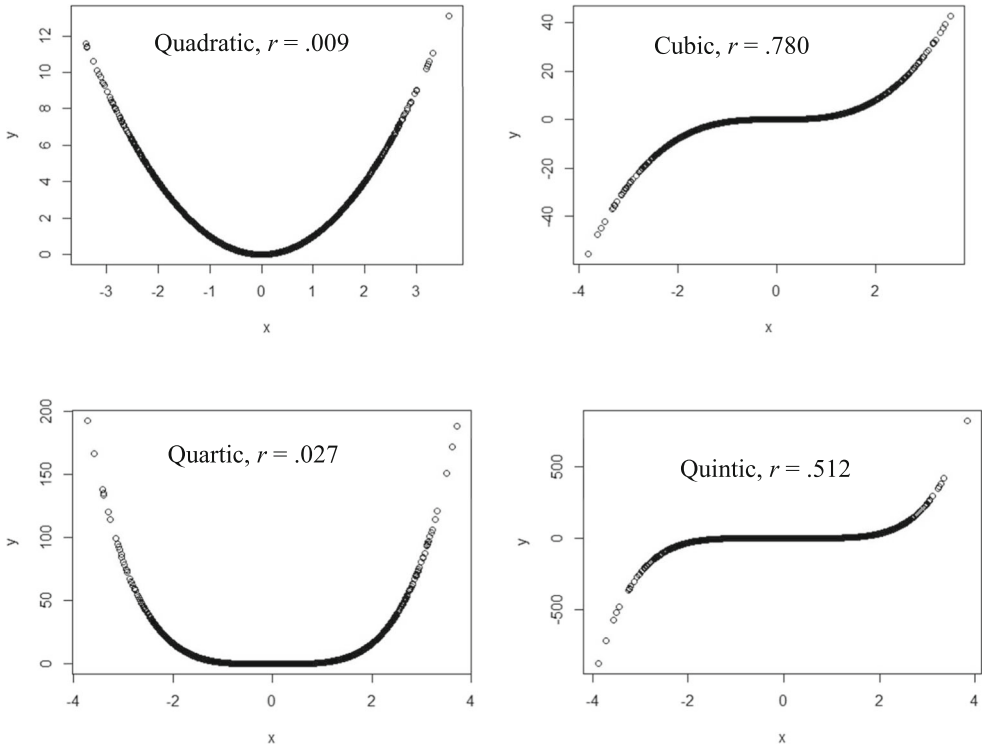

Fig. 2 Scatterplots and observed correlation coefficient *r*s for quadratic, cubic, quartic, and quintic bivariate *X* and *Y*

relationship, making it difficult to evaluate the importance of the relationship. Second, inferences based on this simplified interpretation remain subject to error when the true relationship is not linear. For example, when *X* and *Y* are related by a quadratic function the simplified interpretation of *r* can lead to the incorrect inference that *Y* neither increases nor decreases as *X* increases, when in fact it does both.

Monotonicity measured by nonparametric correlations

Commonly used nonparametric alternatives to *r* are Spearman's rho, Kendall's tau, and robust regression. These alternatives depend on monotonicity, but not linearity, and can still be interpreted as estimates of correspondence. In the case of Spearman's rho, the coefficient of determination describes the proportion of variance in ranks of *Y* scores explained by variance in ranks of *X* scores. Kendall's tau measures the number of concordant *X*–*Y* pairs relative to the number of discordant pairs. Robust regression provides a robust correlation estimate derived from the slope estimated by fitting a robust regression model between *X* and *Y*. The correspondence described by *r* is between scores of *X* and *Y*, whereas the nonparametric alternatives describe correspondence of a different nature, complicating interpretation. Nevertheless, researchers may be tempted to interpret the nonparametric statistics as describing a linear relationship, even if it is not between the scores themselves, an error that could be just as misleading as interpretation of *r* when a linear relationship does not exist (Wilcox, 2012).

PBS measured by common-language effect sizes (CLESs)

Parametric CLES Wolfe and Hogg (1971, p. 30) observed that probability estimates are statistics that "frequently make more sense to the consumers of statistical studies than do the statistics that are now reported in the literature." Their leading example was the probability that an *X* score is greater than a *Y* score. On the basis of this work, McGraw and Wong (1992) proposed that this probability be formalized as a common-language effect size (*CL*). Let $\{X_i \sim N(\mu_i, \sigma_i^2);\ i = 1, 2\}$ be jointly normally and independently distributed random variables that represent responses to two conditions (e.g., treatment and control). McGraw and Wong's *CL* is the sample estimator for $P(X_1 > X_2)$. In equation,

$$CL = \Phi\left[\left(\overline{X}_1 - \overline{X}_2\right)/\sqrt{s_1^2 + s_2^2}\right], \tag{2}$$

where $\left(\overline{X}_1 - \overline{X}_2\right)$ is the sample mean difference, s_i^2 is the sample variance for Group $i = 1, 2$, and Φ is the standard normal distribution function. Simply put, *CL* describes the estimated probability that a score sampled at random from distribution 1

will be larger than a score sampled at random from distribution 2. For example, there is a 70% likelihood that a randomly selected treatment group participant performs better on a cognitive ability test than a randomly selected control group participant. A *CL* value of .5 indicates stochastic equivalence between the two distributions. A value of 1 implies perfect stochastic superiority of one distribution over another. Grissom (1994) derived additional techniques to estimate $P(X_1 > X_2)$ under various conditions and adopted a fitting label to describe this likelihood: probability of superiority.

Nonparametric CLES Vargha and Delaney (2000), expanding on work pioneered by Cliff (1993), later developed a robust estimator (*A*) of *CL*. This development enabled application of the useful and understandable common-language probability-of-superiority conceptualization to data that do not meet parametric assumptions:

$$A = \left\{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\#[y_1(i) > y_2(j)] + .5\#[y_1(i) = y_2(j)]\right\}/n_1 n_2, \quad (3)$$

where # is the count function, $y_1(i)$ and $y_2(j)$ are the *i*th and *j*th observations of *Y* in Groups 1 and 2, respectively, and n_i is the size for Group $i = 1, 2$.

CL and its derivatives are becoming more widely employed (Brooks et al., 2014; Cliff, 1993; Li, 2015; Ruscio, 2008) in behavioral and social sciences research situations involving one nominal variable and one dependent variable. In particular, *A* has been identified as especially useful because it is robust to violations of parametric conditions assumed in *CL* (Ruscio, 2008) and it exhibits characteristics May (2004) has identified as important for effective dissemination of statistical information: understandability, interpretability, and comparability.

A CLES for continuous bivariate data Recognizing that understandability and interpretability were lacking in *r*, Dunlap (1994) proposed an extension of the common-language conceptualization of effect size to research scenarios involving two linearly related bivariate normal variables (i.e., the case in *r*). Dunlap's proposal utilized Sheppard's theorem (Kendall & Stuart, 1977) to convert *r* to a common-language effect size estimate,

$$CL_r = \left[sin^{-1}(r)/\pi\right] + .5, \quad (4)$$

where $\sin^{-1}$ is the inverse sine function and π is a constant ($\approx$ 3.14159; for a mathematical proof, see the Appendix). For example, instead of saying that 16% ($r^2 = .16$) of variance of sons' heights is explained by variance in a 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). Following Grissom's (1994) lead, we describe this likelihood as the probability of bivariate superiority (PBS) and label the resulting parameter γ. We can formalize Dunlap's (1994) conception of CL_r and of PBS as

$$\gamma = P\left(Y > \overline{Y} \cap X > \overline{X}\right) \quad (5)$$

where $\cap$ refers to the intersect function, meaning that both the conditions of $Y > \overline{Y}$ and $X > \overline{X}$ should be met for the joint probability of γ. In practice, researchers may not have a full dataset from the entire population, and B_p is denoted as a sample estimator for γ.

Understanding a bivariate relationship in terms of γ is conceptually similar to Blomqvist's (1950) q' test of dependence. Whereas γ is concerned with the distribution of *XY* scores evaluated with reference to the mean values of *X* and *Y*, Blomqvist's procedure plots the bivariate data into four quadrants according to the median values of *X* and *Y*. The q' test is based on the count of scores in each quadrant and the resulting metric lies between – 1 and + 1, which brings with it interpretive difficulties similar to *r*. When the condition of bivariate normality is met, the mean and median of *X* are equal, and the mean and median of *Y* are equal. Under these conditions CL_r is mathematically equivalent to *q'* (see the Appendix) but is expressed as a likelihood for greater interpretability.

It is apparent to us that Dunlap (1994) intended his extension of *CL to* CL_r chiefly to improve understandability of linear relationships by describing them in intuitive terms. His work was a worthwhile undertaking and an impressive innovation that opened the door to a new potential understanding of bivariate relationships. CL_r is a more understandable way to describe the relationship between *X* and *Y*, as a probability of superiority, and it makes both the direction and the magnitude of the relationship comprehensible. But CL_r, like q', describes a different relationship than *r*, and does so in terms that are unrelated to linear correspondence. In fact, reading Dunlap's description and looking at Eq. 5, it is apparent that γ, and therefore CL_r, describes a relationship that is not necessarily linear. Nevertheless, Dunlap's conversion formula is based on the assumption that *X* and *Y* are linearly related, and this limits its usefulness.

Equation 4 implies a dependence between the existence of linearity and the existence of a probability-of-superiority effect. However, this implication may not hold. It is possible for a probability-of-superiority relationship to exist between *X* and *Y* in the absence of a linear relationship. Figure 3 depicts two idealized plots of the probability of bivariate superiority. Plot 3.A.vii shows a perfect linear relationship between *X* and *Y* ($r = 1$), which implies that a PBS relationship exists, with $B_p = CL_r = 1$. This example is congruent with Dunlap's (1994) assertion that linearity is sufficient for a PBS relationship to exist. Plot 3.B.vii depicts a perfect PBS relationship ($B_p = 1$) in which the underlying

relationship between X and Y is not linear. This example shows that a linear relationship is not a necessary condition for a PBS relationship to exist.

Since Dunlap's (1994) bold advance, there has been no development of CL_r as a measure of probability of superiority in bivariate relationships independent from r. It is worth speculating that Dunlap's conversion formula could be applied to nonparametric correlation coefficients when the parametric assumptions for r are violated, but this possibility has not previously been investigated. We tested the usefulness of such an application in our Monte Carlo experiment. Our central contribution is the introduction of a method to directly estimate PBS in the absence of a linear relationship.

Our proposed B_p: A nonparametric extension of CL_r

Taking inspiration from the work of Vargha and Delaney (2000), who developed the robust estimator A for CL, we have developed a robust estimator, B_p, of Dunlap's (1994) CL_r. B_p estimates the magnitude of a PBS relationship between two variables, X and Y, without the restrictions of bivariate normal correlation that are required for the conversion in Eq. 4. Like CL_r, B_p is conceptualized as the probability that when an X score is above (or below) the mean of all X scores, its paired Y score is also above (or below) the mean of all Y scores, as formalized in Eq. 5. When X and Y follow a bivariate normal distribution and form a linear relationship, B_p directly estimates CL_r without relying on a transformation from r.

It is noteworthy that PBS is a special case under copula theory. A copula is used to measure joint distributions of two or more random variables (e.g., Botev, 2017; Jaworski, Durante, Härdle, & Rychlik, 2010; Nelson, 2006). A review of copula theory as it relates to bivariate relationships is beyond the scope of this article, so we refer the reader to Lai and Balakrishnan (2009, chap. 2) for a review. In lay terms, a copula is a statistical concept that explains how two variables are related to each other. In the bivariate X–Y case, a copula allows one to separate the joint X–Y distribution into two sources: the marginal distributions of each variable, and the copula that "glues" these variables into together. Linear association is only one example of many different types of "glues." As we noted above, limiting investigation of bivariate relationships to the linear copula restricts the potential to identify other bivariate relationship forms that may be practically and theoretically important. Researchers need accessible methods to examine relationships described by a wide array of copulas, including PBS.

Applying copula theory to Blomqvist's (1950) Eq. 2, we can define how the distribution of Y is glued to the distribution of X. Replacing Blomqvist's medians with means in our division of quadrants, the copula determines the probability that an XY point falls within a particular quadrant given a value of γ. Let X_i follow a probability distribution (e.g., normal, lognormal, uniform, etc.). There exists a marginal probability distribution for Y_i that is generated from the following function:

$$\begin{aligned}
&Y_i \sim U(\mu_Y, c), \text{if } X_i > \mu_X \text{ and } \varrho \leq \gamma,\\
&Y_i \sim U(\mu_Y, c), \text{if } X_i > \mu_X \text{ and } \varrho > \gamma,\\
&Y_i \sim U(-c, \mu_Y), \text{if } X_i < \mu_X \text{ and } \varrho \leq \gamma,\\
&Y_i \sim U(-c, \mu_Y), \text{if } X_i > \mu_X \text{ and } \varrho > \gamma,\\
&Y_i = \mu_Y, \text{if } X_i = \mu_X,
\end{aligned} \tag{6}$$

where c is the limit in a uniform distribution, μ_X is the population mean of X, μ_Y is the population mean of Y, $\varrho \sim U(0, 1)$ follows a uniform distribution with min = 0 and max = 1, and γ is the population PBS that relates X and Y. The generated ϱ values control the likelihood that a simulated Y score is above (or below) the mean of Y, when a simulated X score is above (or below) the mean of X, so that Y is related to X for a particular value of γ.

When the condition of bivariate normality is met, a correlational estimate (r) can be translated into the more understandable PBS using Dunlap's (1994) common-language correlation transformation (CL_r). To continue with the example in our introduction above, if data for physical activity and mortality meet the bivariate normality condition, then an estimated r of .60 can be converted to the PBS estimate of .705. Below we describe the algorithm by which this estimate can be directly computed without reference to r. Use of this algorithm makes it possible to directly compute the B_p estimate and detect PBS-based relationships even when conversion using CL_r is also inappropriate, such as when bivariate normality is violated. Thus, B_p provides a robust estimator of the PBS-based relationship in the population, which we label γ.

Figure 3 shows scatterplots for X and Y when (A) X and Y are generated from the conventional condition of linearly related X and Y that forms the bivariate normal correlation (such that r can be mathematically linked to CL_r in Eq. 4, which estimates the probability in Eq. 5), and (B) X and Y are directly generated from the PBS function (i.e., Eq. 6) without the unnecessary condition of linearity. In other words, the plots in column B show the relationship between X and Y is based only on a level of γ, and the plots in column A demonstrate the same type of PBS-based relationship between X and Y when the condition of linearity is also met. Whereas Dunlap's (1994) CL_r can accurately detect PBS if X and Y are linearly related and follow bivariate normal distributions, PBS-based relationships can exist when these conditions are not met.

As was noted by Reshef et al. (2011), a good statistic for measuring dependence should possess two heuristic properties—generality and equitability. *Generality* means that a statistic should detect and measure a wide range of possible associations between X and Y, not limited to specific relationships (e.g., linear). *Equitability* means that a statistic should give similar values "to equally noisy relationships of any types" (p. 1518). In light of these properties, we have developed a statistic (B_p; Eq. 7 below) for estimating the population PBS (γ) such that it can detect and estimate PBS whether or not a linear X–Y relationship exists.

First, we need an algorithm that can measure the number of times that Y is above (or blow) the mean of Y when the corresponding X score is above (or below) the mean of X. The count function, $\#[\text{sign}\,(x_i-\bar{x})\cdot\text{sign}\,(y_i-\bar{y}) > 0]$, serves this purpose. Second, there is a scaling algorithm, $0.5\#[\text{sign}\,(x_i-\bar{x})\cdot\text{sign}\,(y_i-\bar{y}) = 0]$; when $(x_i-\bar{x})$ and/or $(y_i-\bar{y})$ equal to 0, then a .5 unit is assigned to the B_p calculation. The purpose of this algorithm is to ensure that when there is zero PBS relationship for all the X and Y scores, the summation of the count will become half of the sample size (n), and hence, the B_p score (Eq. 7) is scaled to become .50, as in the case of the calculation for the A statistic in Eq. 3.

Given these considerations, we derived

$$B_P = \frac{\sum_{i=1}^{n}\#[\text{sign}\,(x_i-\bar{x})\cdot\text{sign}\,(y_i-\bar{y}) > 0] + 0.5\#[\text{sign}\,(x_i-\bar{x})\cdot\text{sign}\,(y_i-\bar{y}) = 0]}{n}, \tag{7}$$

where n is the size or number of paired observations, # is the count function, x_i and y_i are the scores or observations from an X–Y pair in a sample, $\bar{x}$ is the sample mean of X, and $\bar{y}$ is the sample mean of Y.

We expect that B_P can give similar PBS scores for both the linear-based and PBS-based X–Y planes in Fig. 3, which meet the scientific features of generality and equitability for a good statistic (for details about the mathematical relationship between B_p and CL_r; please see the Appendix). We conducted a Monte Carlo experiment to evaluate the behavior of B_P and compared performance of this new statistic to the performance of Dunlap's conversion formula applied to r and its nonparametric counterparts.

It is noteworthy that, even though B_P appears to be and is positioned as a more general, interpretable, and robust statistic for measuring PBS-based bivariate relationship, it is not the most powerful statistic for detecting monotonic or linear relationships. Nor is B_P an alternative to correlation coefficient r and nonparametric correlations (e.g., rank correlation) when a linear or monotonic relationship is hypothesized. In other words, the correlation coefficient r works best when estimating the linear relationship, $Y = aX + b$, and nonparametric correlations (e.g., rank correlation) are expected to work well when estimating linearity and monotonicity. In particular cases in which the underlying relationship between X and Y is monotonic, the nonparametric correlations (e.g., rank correlation) will generally indicate a stronger relationship than B_P because they use more information about the data than just relationship to the mean. However, the rank correlation may fail when the relationship is no longer monotonic, as in the data following Eq. 6. Specifically, given that Eq. 6 sets up data to minimize the relationship between X and Y beyond the PBS relationship, X and Y are basically unrelated within a quadrant and particularly are not monotonically related within a quadrant. In short, despite previous research findings about the improved interpretability of PBS-related statistics, such as CL_r, researchers should consider the type of bivariate relationships (i.e., linearity, monotonicity, and PBS) they are focusing on, and choose the corresponding statistic (e.g., r, rank correlation, or B_P).[2]

Design of the Monte Carlo experiment

Comparative estimates

Numerous robust estimators have been proposed for detecting X–Y associations when the parametric assumptions for r have been violated. Three common robust correlations, Spearman's rank correlation (r_s; Kendall & Gibbons, 1990), Kendall's tau correlation (r_t; Kendall, 1938), and robust regression correlation (r_r; Wilcox, 2012), can be converted to a CL_r value using Eq. 4. For comparative purposes this study examines the performance of such conversion of these robust correlation estimators, as well as the original Pearson-based CL_r as defined by Dunlap (1994), as estimators of γ.

Spearman-based estimation (CL_S) Spearman's correlation r_s converts X and Y scores into rank scores, then applies Pearson's product-moment correlation formula to calculate the distance between these rank scores and summarize an overall relationship between the ranks of X and Y scores. The resulting value can be plugged into (Eq. 4) to obtain an estimate of γ,

$$\begin{aligned} CL_S &= \left[sin^{-1}(r_s)/\pi\right] + .5 \\ &= \left\{sin^{-1}\left\{1-6\textstyle\sum_{i=1}^{n} d_i^2/\left[n\left(n^2+1\right)\right]\right\}/\pi\right\} + .5, \end{aligned} \tag{8}$$

[2] We are grateful to an anonymous reviewer for reminding us of the important implications for hypothesis testing of the differentiation between linearity, monotonicity, and PBS.

where $d_i^2 = [\text{rank}(x_i) - \text{rank}(y_i)]^2$ is the squared difference between the rank of a score in X and a score in Y.

Kendall-based estimation (CL_T) Kendall's tau (r_t) measures the strength of association between X and Y, which can be plugged into Eq. 4 to obtain an estimate of γ,

$$CL_T = \left[\sin^{-1}(r_t)/\pi\right] + .5 = \left\{\sin^{-1}\{(n_C - n_D)/[0.5 \cdot n(n-1)]\}/\pi\right\} + .5, \quad (9)$$

where n_C refers to the number of concordant pairs for X and Y, and n_D refers to the number of discordant pairs for X and Y.

Regression-based estimation (CL_L) To obtain a robust correlation (r_r), one can fit a regression line that regresses the standardized Y scores (Z_Y) on the standardized X scores (Z_X) based on a robust estimation procedure (e.g., M-estimation; Wilcox, 2012)—that is,

$$Z_Y = \beta_0 + \beta_1 Z_X + e, \quad (10)$$

where the standardized slope β_1 denotes a robust correlation between X and Y that can be plugged into (4) to obtain an estimate of τ,

$$CL_L = sin^{-1}(r_r)/\text{pi} + .5 = sin^{-1}(\beta_1)/\pi + .5. \quad (11)$$

We hereafter refer to the various estimates of γ as PBS = (B_P, CL_r, CL_S, CL_T, and CL_L) in general.

Bootstrap confidence intervals

In addition to the point estimate of γ, its confidence interval (CI) is essential for quantifying the sampling error and making statistical inferences. For example, if a 95% CI for B_p does not span .50 it can be inferred that the PBS estimate is statistically significant at the .05 level. Bootstrapping (Efron & Tibishiri, 1993)—a nonparametric resampling procedure often executed in a computerized statistical package—can often produce trustworthy CIs for statistical measures (Chan & Chan, 2004; Li, Chan, & Cui, 2011). There are three major types of bootstrap CIs: bootstrap standard interval (BSI), bootstrap percentile interval (BPI), and bootstrap bias-corrected and accelerated percentile interval (BCaI).

Assume one has a dataset with 100 X–Y paired observations. First, this dataset is resampled with replacement B times (e.g., 1,000) to produce N bootstrap datasets that contains the same sample size (i.e., 100) as the original dataset. Second, for each of the N = 1,000 bootstrap datasets, the PBS point estimates (denoted as PS in Equations 12 to 16) are computed using a statistical package, thereby producing 1,000 bootstrap $PS^* = B_P^*(b)$, $CL_r^*(b)$, $CL_S^*(b)$, $CL_T^*(b)$, and $CL_L^*(b)$ estimates, where $b = 1, 2, \ldots, N$. Given these bootstrap estimates, the statistical package can construct the 95% BSI,

$$\text{BSI} = \widehat{PS} \pm 1.96 \cdot s_{PS}^*, \quad (12)$$

where $\hat{PS}$ is an estimated B_P, CL_r, CL_S, CL_T, and CL_L respectively, from the original dataset, and s_{PS}^* refers to the standard error for B_P, CL_r, CL_S, CL_T, and CL_L respectively, on the basis of the standard deviation of the 1,000 bootstrap samples.

A second method is known as the 95% BPI,

$$\text{BPI} = \left(PS^*(l), PS^*(u)\right), \quad (13)$$

where l is the 2.5 percentile rank and u is the 97.5 percentile rank of the 1,000 bootstrap $PS^* = B_P^*(b)$, $CL_r^*(b)$, $CL_S^*(b)$, $CL_T^*(b)$, and $CL_L^*(b)$ estimates, respectively.

A third type of bootstrap CI is the 95% BCaI that adjusts for any skewness in the original dataset,

$$\text{BCaI} = \left(PS^*\left(l^*\right), PS^*\left(u^*\right)\right), \quad (14)$$

where the l^* and u^* are lower and upper percentile ranks, which are adjusted to be different from l = 2.5% and u = 97.5% in Eq. 14, depending on the skewed level of the original dataset. Two correction factors, i, and j, are required to estimate l^* and u^*. The first factor i, is used to correct for the overall bias of the bootstrap PS^* estimates, which deviate from the estimate obtained in the original dataset. That is,

$$i = \Phi^{-1}\left\{\#\left[PS^*(b) < PS\right]/N\right\}, \quad (15)$$

where Φ^{-1} is normal inverse cumulative function distribution, and $\#[PS^*(b) < PS]$ is the count function that counts the number of the bootstrap PS^* estimates smaller than PS in the original dataset. The second factor (j) adjusts for the rate of change of the error of PS with respect to its true parameter value,

$$j = \sum_{k=1}^{K}[PS(.) - PS(k)]^3 / 6\left\{\sum_{k=1}^{K}\left[PS(.) - PS(k)^2\right]^{3/2}\right\}, \quad (16)$$

where $PS(k)$ is the jackknife value of PS obtained by removing the kth row of the original dataset, and $PS(.)$ is the mean of the n jackknife estimates. Consequently, $l^* = N \cdot \alpha_1$, where $\alpha_1 = \Phi\left\{i + \frac{i + z_{1-(\alpha/2)}}{1 - j\left[i + z_{1-(\alpha/2)}\right]}\right\}$, and $u^* = N \cdot \alpha_2$, where $\alpha_2 = \Phi\left\{i + \frac{i - z_{1-(\alpha/2)}}{1 - j\left[i - z_{1-(\alpha/2)}\right]}\right\}$ (for details, please see Canty & Ripley, 2016; Chan & Chan, 2004; Efron & Tibshirani, 1993; Li et al., 2011)

Simulation conditions

Our experiment investigated the performance of PBS estimators under three bivariate relationship conditions—(1) linearly related X and Y that follow bivariate normal correlation, (2) linearly related X and Y that follow nonnormal correlations (i.e., positively skewed, negatively skewed, and uniform), and (3) PBS related X and Y that follow normal, uniform, positively skewed, and negatively skewed distributions. For each of these relationships, seven levels of true effect size (γ), and four levels of sample size were evaluated.

Population effect size (seven levels) Seven effect size levels were evaluated: $\gamma = .50, .55, .60, .65, .70, .75$, and .80. When assumptions for r are met, these values can be converted using Eq. 4 to the ρ values 0, .156, .309, .454, .588, .707, and .809, providing a comprehensive span of effect sizes [zero, small (.1), medium (.3), large (.5), and extremely large (.8); Cohen, 1988].

Sample size (four levels) Four levels, $n = 20$, 60, 100, and 300, were evaluated to represent a range of common sample sizes in behavioral and social science research.

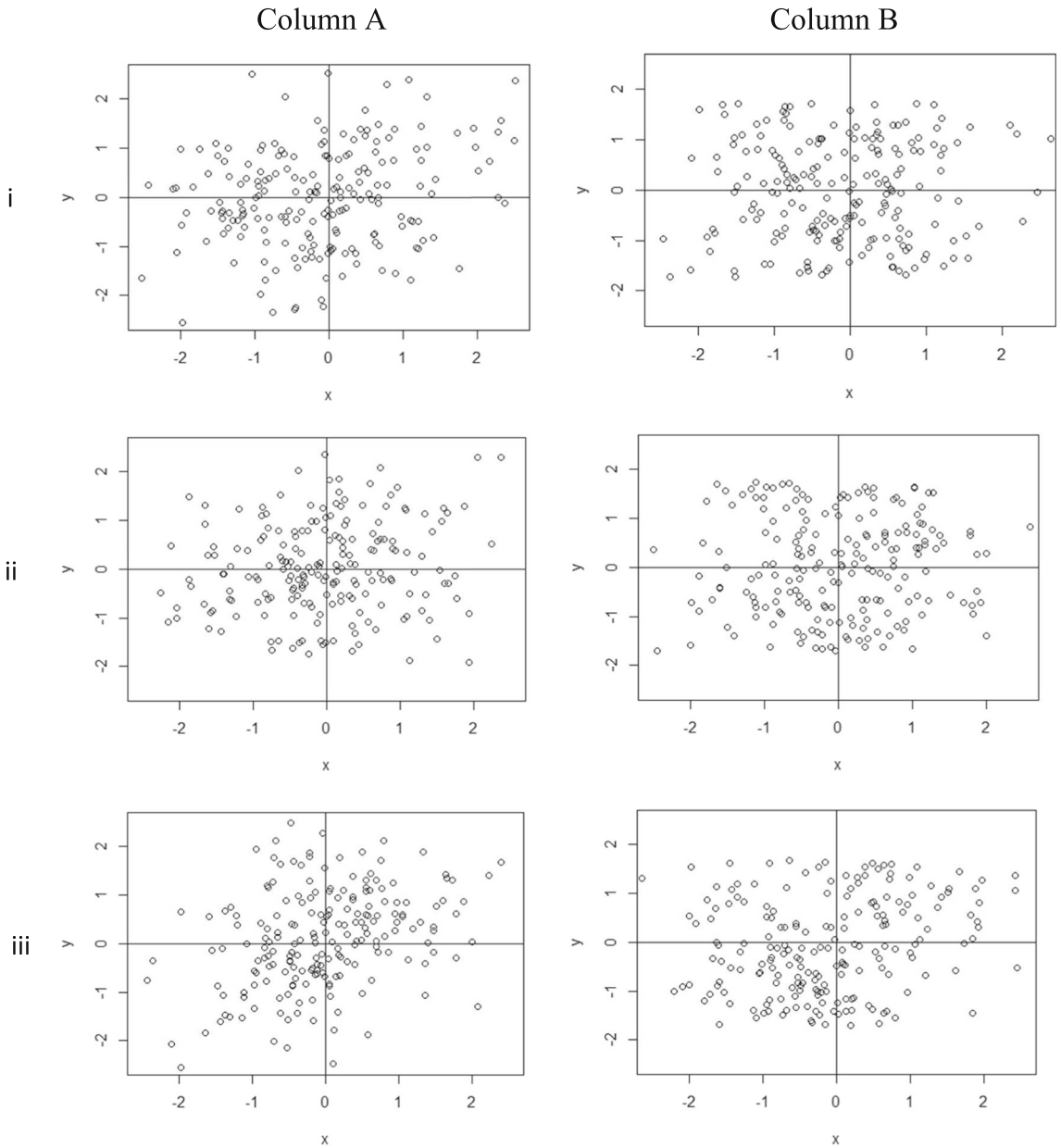

Fig. 3 Sample scatterplots for bivariate normal correlations with $\rho = .05$, .1, .3, .5, .7, .9, and 1 (left) (which, when transformed using Eq. 4, provides $CL_r = .516$, .532, .597, .667, .747, .856, and 1), as well as the equivalent probability-of-bivariate-superiority relationships without the requirement of a linear relationship, with the true PBS values $\gamma = .516$, .532, .597, .667, .747, .856, and 1 (right)

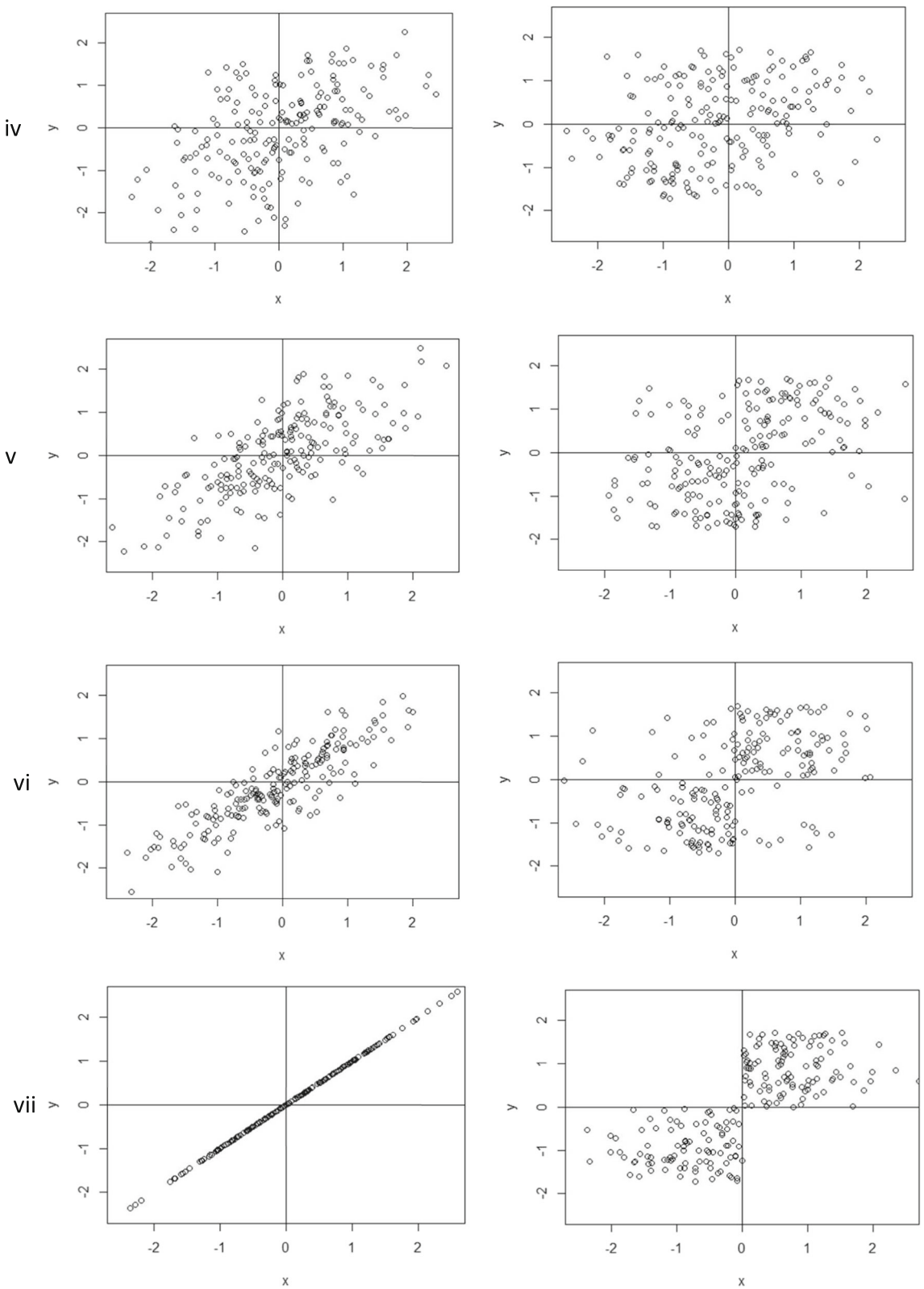

Fig. 3 (continued)

Bivariate Type 1 distributions (four levels) Type 1 is separated into Type 1a (bivariate normal and linear) and 1b (bivariate nonnormal and linear). Type 1a data meet both the PBS-based condition for B_p plus the additional linear condition for r and CL_r. In other words, Type 1a meets the parametric assumptions that satisfy both r and PBS. Type 1b data are uniform, positively

skewed, or negatively skewed and meet the linear condition for X and Y but do not necessarily meet the PBS condition for X and Y. The X scores were simulated to behave different from normality, and Y scores simulated conditional on the X scores for a level of ρ. We expected that B_p and CL_r would behave differently, because the necessary condition (bivariate normal correlation) that links B_p to CL_r was violated.

Bivariate Type 2 distributions (4 × 3 levels) The aforementioned distributions do not allow for a full demonstration of potential PBS-based relationships, as they are based on the widely employed concept of linearly related X and Y. An important gap in previous research is that PBS-based relationships that are free of linearity have been ignored. These bivariate relationships can be directly derived from the function in Eq. 6. Given this function, X can be generated from any distributions (e.g., normal, uniform, positive skewed, negative skewed), and Y can be generated from Eq. 6 with a manipulated level of γ. In this study, three levels of c—$\sqrt{3}/2$, $\sqrt{12}/2$, and $\sqrt{48}/2$ —in Eq. 6 were examined, which produced three types of uniformly distributed Y scores with SDs = 0.25, 1, and 4, which have a smaller, identical, and larger SD than the SD (i.e., 1) of the X scores. For SDs of Y equal to either .25 or 4, the generated X scores appear to contain outliers. These distributions reflect scenarios in which some variables contain larger variance and have longer tails than others.

This experiment was designed with 7 × 4 × 4 = 112 simulation conditions that meet the linear condition and 7 × 4 × 4 × 3 = 336 simulation conditions that meet only the PBS conditions for a total of 448 simulation conditions. Each condition was replicated 1,000 times, and for each replication the dataset was bootstrapped 1,000 times to generate the three bootstrap CIs. This produced a total of 448(conditions) × 1, 000(replication) × 1, 000(bootstrap) = 448,000,000 simulated datasets.

Procedure

Type 1 To generate the simulation data, first, X scores were generated from a normal distribution, $N(0, 1^2)$, which meets the linear condition. Second, X scores were generated from a uniform distribution, $U(-\sqrt{12}/2, \sqrt{12}/2)$, with mean = 0, SD = 1; this mimics scenarios in which scores are evenly and uniformly distributed in a sample. Third, X scores were generated from a lognormal distribution, $\ln N(-0.3456, 0.8326^2)$, so that the mean is 1 and SD is 1. This forms a positively skewed distribution, with skewness = 4 and kurtosis = 38, commonly found in behavioral and social sciences, for example, in data from biological measures (e.g., Wilcox, Granger, Szanton, & Clark, 2014) and measures of affect and depression levels (e.g., Tomitaka et al., 2016). Fourth, X scores were generated on the basis of – 1 multiplied by X_a scores, which were generated from $\ln N(-0.3456, 0.8326^2)$. Hence, the generated X scores follow a negatively skewed distribution with mean = – 1, SD = 1, skewness = – 4, and kurtosis = 38. Given the generated X scores, for Type 1, the linear-related Y scores were generated from

$$Y = \rho X + e_Y, \tag{17}$$

where ρ is the population Pearson's correlation converted from the population PBS (γ) through Eq. (4), X refers to the simulated scores from Type 1 or 2, and e_Y is the error score generated from a normal distribution with mean = 0, and variance = $1 - \rho^2$. Given this method, X and Y are expected to be linearly correlated with a level of ρ.

For Type 2, according to Eq. 6, ϱ values were generated from a uniform distribution, U(0, 1), with min = 0 and max = 1. Second, to allow sampling error in each replicated sample, the γ values were generated from a binomial distribution, $B(n, \gamma)$. Given the generated ϱ and γ for each simulated respondent, and the known X scores, the Y scores were generated from either a uniform distribution [i.e., $U(\overline{X}, c), U(-c, \overline{X})$] or set at the mean of Y as shown in Eq. 6. Note that to allow for sampling error, the sample mean estimate $\overline{X}$ was used instead of μ_X in Eq. 6.

Consequently, for each replication, a dataset was generated containing both the X and Y scores that were used to compute the PBS estimates of γ. This dataset was also used to generate the 95% BSI, BPI, and BCaI (with 1,000 bootstrap replications). Thus, for each condition 1,000 PBS = (B_P, CL_r, CL_S, CL_T, and CL_L) estimates and their associated 95% BSI, BPI, and BCaI were obtained. The simulation was conducted in R (R Core Team, 2016). Note that the code called two packages, boot (Canty & Ripley, 2016) and MASS (Venables & Ripley, 2002), that executed the bootstrap procedures and performed the robust linear regression, respectively. The simulation code is available in the supplementary materials.

Results

Evaluation criteria

For Distribution Types 1a and 2, two evaluation criteria are used to assess the performance of each of the PBS estimates of γ. First, percentage bias was computed as bias = $[(\overline{PS}-\gamma)/\gamma]$, where $\overline{PS}$ is the mean of the 1,000 PBS estimates (expressed as B_P, CL_r, CL_S, CL_T, and CL_L) in 1,000 simulated samples. A PBS estimate was considered reasonable if the bias was within ± .10 (or 10%; Li et al., 2011). This bias only examines the performance of a PBS estimate in one condition. To evaluate overall performance across all 336 nonparametric conditions a second criterion was used: the mean-absolute percentage bias (MAPE) is defined as MAPE = $\left(\sum_{s=1}^{336}|\text{bias}(s)|\right)/336$. A MAPE smaller than .10 (or 10%) is considered

reasonable (Li et al., 2011). MAPE was also used to separately evaluate the performance of the 28 Type 1a linear and normal conditions, in which MAPE $= \left(\sum_{s=1}^{28}|\text{bias}(s)|\right)/28.$, and the 84 Type 1b linear and nonnormal conditions, in which MAPE $= \left(\sum_{s=1}^{84}|\text{bias}(s)|\right)/84$. Regarding the performance of bootstrap CIs, given that 95% BSI, BPI, and BCaI were constructed, coverage was expected to be 950 out of 1,000 replications [or expressed as coverage probability (CP) = .95]. But sampling error makes it impractical for researchers to obtain a perfect CP of .95. We considered acceptable an observed CP that falls within the range (.925, .975) (Chan & Chan, 2004).

For Distribution Type 1b, the purpose is to evaluate the difference between B_p and CL_r (i.e., $=B_p - CL_r$) when X–Y points were generated from nonnormal distributions (i.e., positively skewed, negatively skewed, uniform) with an associated true correlation value (ρ). Given that X was generated from a nonnormal distribution, and Y was generated from a linear model (Eq. 17) conditional on a true correlation value (ρ) related to X, the equivalence for $B_p \equiv CL_r$ [i.e., $B_P = \frac{\sum_{i=1}^{n}[\text{sign}(x_i-\bar{x})\cdot\text{sign}(y_i-\bar{y})>0]+0.5[\text{sign}(x_i-\bar{x})\cdot\text{sign}(y_i-\bar{y})=0]}{n} \equiv \frac{1}{\pi}\sin^{-1}(r) + 0.5 = CL_R$; Eq. 8 in the Appx.] becomes invalid. Hence, even though we know the true ρ value in our simulation, the associated true PBS value, $\gamma = P\left(Y > \overline{Y} \middle| X > \overline{X}\right)$, is unknown and is unique for every type of nonnormal distribution of X. As a result, it is inaccurate for us to evaluate biases and coverage probabilities when γ is unknown.

Performance under linear conditions (Table 1)

Type 1a As expected, across the 28 conditions in which linear and normal conditions were met, the performances of B_P and CL_r were highly comparable. The biases of B_P ranged from – .005 to .008, with mean .000 and *SD* .003, indicating an excellent fit. Of the 28 conditions, all biases fell within the criterion of ± .10 (or 10%), and MAPE was .002. For CL_r, the biases ranged from – .007 to .004, with mean .000 and *SD* .002, showing an excellent fit. All the biases were within the criterion of ± .10 (or 10%), and MAPE was .001. Regarding the bootstrap CIs, the mean of the 28 CPs yielded by BSI, BPI, and BCaI for B_P were .966, .978, and .913, respectively, which are comparable to those obtained by BSI, BPI, and BCaI for CL_r, i.e., .932, .937, and .942, respectively. Thus, both the new PBS procedure (producing the B_P estimate) and the traditional CL_r are trustworthy and appropriate when the parametric assumptions are met: Both methods produced similar results with minimal bias.

Type 1b As is shown in Fig. 4, B_P consistently produced a larger estimate than CL_r under a uniform distribution: The difference (*d*) values ranged from .000 to .034, with a mean of .015. For positively and negatively skewed distributions, B_P consistently produced a smaller estimate than CL_r: Here the *d* values ranged from – .046 to .001, with a mean of – .022. These results indicate that B_P and CL_r are different—that is, $B_P = \frac{\sum_{i=1}^{n}\#[\text{sign}(x_i-\bar{x})\cdot\text{sign}(y_i-\bar{y})>0]+0.5\#[\text{sign}(x_i-\bar{x})\cdot\text{sign}(y_i-\bar{y})=0]}{n} \neq \frac{1}{\pi}\sin^{-1}(r) + 0.5 = CL_R$, when the linear and nonnormal conditions are met.

Table 1 Coverage probabilities (CP) and percentage biases of the point estimates of B_P and, CL_r when the linear and normal conditions are met (Type 1)

		B_P				CL_r			
			CP				CP		
γ	n	% Bias	BSI	BPI	BCaI	% Bias	BSI	BPI	BCaI
.50	20	.008	.972	**.988**	**.903**	.001	**.912**	.929	.948
	60	.001	.956	.969	**.901**	.001	**.909**	**.922**	.927
	100	.005	.958	.967	**.923**	.001	.929	.930	.932
	300	.001	.967	.969	.941	.001	.949	.950	.955
.55	20	– .004	.968	**.983**	**.893**	– .007	**.906**	**.920**	.934
	60	.001	.966	.975	**.924**	.004	.940	.951	.951
	100	– .002	.971	**.976**	.927	.000	.946	.947	.953
	300	.001	.957	.957	.931	.000	.932	.937	.936
.60	20	– .001	.973	**.987**	**.884**	– .002	**.917**	**.924**	.936
	60	– .002	.963	.973	**.918**	– .002	.935	.939	.937
	100	.002	.966	**.979**	**.924**	.001	.933	.936	.940
	300	– .001	.961	.970	.934	.001	.945	.949	.946
.65	20	– .004	**.978**	**.994**	**.896**	.001	.928	.935	.947
	60	.000	.958	.973	**.900**	.002	.925	.929	.935
	100	.001	.965	.973	.936	.000	.958	.961	.962
	300	– .002	.959	.968	.928	– .001	.963	.963	.963
.70	20	.002	.969	**.988**	**.896**	– .001	**.906**	**.911**	.928
	60	– .004	.963	**.982**	**.892**	– .002	.934	.941	.938
	100	.002	.962	.972	**.922**	.002	.930	.936	.942
	300	.001	.954	.966	.935	.001	.931	.933	.929
.75	20	– .005	.975	**.994**	**.885**	– .002	**.920**	**.919**	.940
	60	.002	.966	**.984**	**.911**	.002	.928	.931	.939
	100	.003	.963	.972	**.917**	.000	.942	.945	.951
	300	– .001	.965	.966	.926	.000	.942	.941	.941
.80	20	.004	.971	**.996**	**.867**	.002	**.896**	**.900**	**.919**
	60	.001	**.976**	**.988**	**.912**	.001	.946	.951	.953
	100	– .002	**.976**	**.987**	**.914**	– .001	.951	.952	.953
	300	– .001	.965	**.977**	.930	.000	.943	.942	.941

γ is the population PBS parameter, n is the sample size, B_P is the proposed robust estimator for the population PBS, CL_r is the Pearson-based estimator for the population PBS, BSI is the 95% bootstrap standard interval, BPI is the 95% bootstrap percentile interval, and BCaI is the 95% bootstrap bias-corrected and accelerated percentile interval. Coverage probabilities outside acceptable range are presented in bold

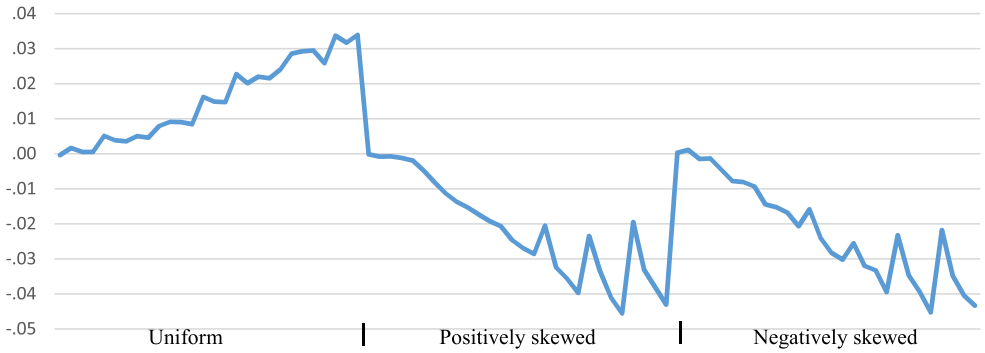

Fig. 4 Means of 1,000 replicated estimates for B_P minus CL_r (i.e., *d* scores) across the 84 conditions in which linear and nonnormal conditions are met (Type 1b).

Performance under PBS conditions in the absence of linearity

Point estimates Among the five *PBS* estimates, B_P performed best and was the only estimate considered reasonable across both evaluation criteria, as is shown in Fig. 5. Across the 336 conditions in which parametric assumptions were violated, the biases ranged from – .059 to .057, with mean – .018 and *SD* .017, indicating a good estimate. All 336 conditions produced a B_P estimate within the criterion of ± .10 (or 10%), showing excellent fit. Overall, the MAPE (.019) was also well within the criterion of ± .10 (or 10%), which demonstrates that B_P is an appropriate and robust estimator of γ across all the simulation conditions.

The performance of the remaining four *PBS* estimates was less than optimal. The parametric-based CL_r performed poorly and did not reliably detect PBS relationships. The biases ranged from – .234 to .007, with mean – .119 and *SD* .070. Of the 336 conditions, only 128 (38.1%) yielded a CL_r estimate within ± .10. Overall, the MAPE for CL_r was outside the criterion (.119), further demonstrating that CL_r is not an optimal estimator for γ when *X* and *Y* are not linearly related.

Two robust-based *PBS* estimates performed slightly better than CL_r but neither performed adequately. The CL_S biases ranged from – .234 to .007, with mean – .113 and *SD* .070. Of the 336 conditions, 144 (or 42.9%) produced a CL_S estimate within ± .10, and the MAPE was .114. The CL_L biases ranged from – .210 to .007, with mean – .110 and *SD* .064. Of the 336 conditions, 139 (or 41.4%) produced a CL_L estimate within ± .10, and the MAPE was .110. The last robust estimate, CL_T, performed worse than the parametric CL_r. These biases ranged from – .267 to .004, with mean – .146 and *SD* .086. Of the 336 conditions, only 96 (or 28.7%) produced a CL_T estimate within ± .10, and the MAPE was .146.

Confidence intervals As is shown in Fig. 6, two of the three bootstrap CI procedures provided acceptable coverage probabilities (CPs) for B_P. The 95% BSI for B_P performed best: Across the 336 conditions, the CPs obtained from the 95% BSI ranged from .666 to .986, with mean .954 and *SD* .043. Of the 336 conditions, 280 (or 83.3%) conditions produced a CP within the criterion of (.925, .975), indicating good fit. The 95% BPI for B_P yielded CPs that ranged from .422 to .998, with mean .934 and *SD* .084. Of the 336 conditions, 229 (or 68.2%) conditions produced a CP within (.925, .975). However, the 95% BCaI for B_P was less than optimal: The CPs ranged from .652 to .964, with mean .901 and *SD* .041. Of the 336 conditions, 109 (or 32.4%) conditions produced a CP within (.925, .975).

Given that the bias of the point estimates based on CL_r, CL_S, CL_T, and CL_L are outside a reasonable range, the associated

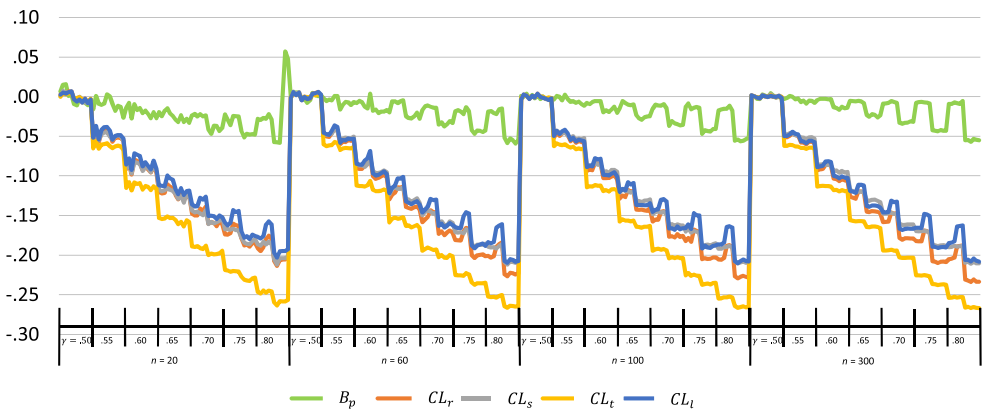

Fig. 5 Biases for B_P, CL_r, CL_S, CL_T, and CL_L when only the PBS condition is met (Type 2)

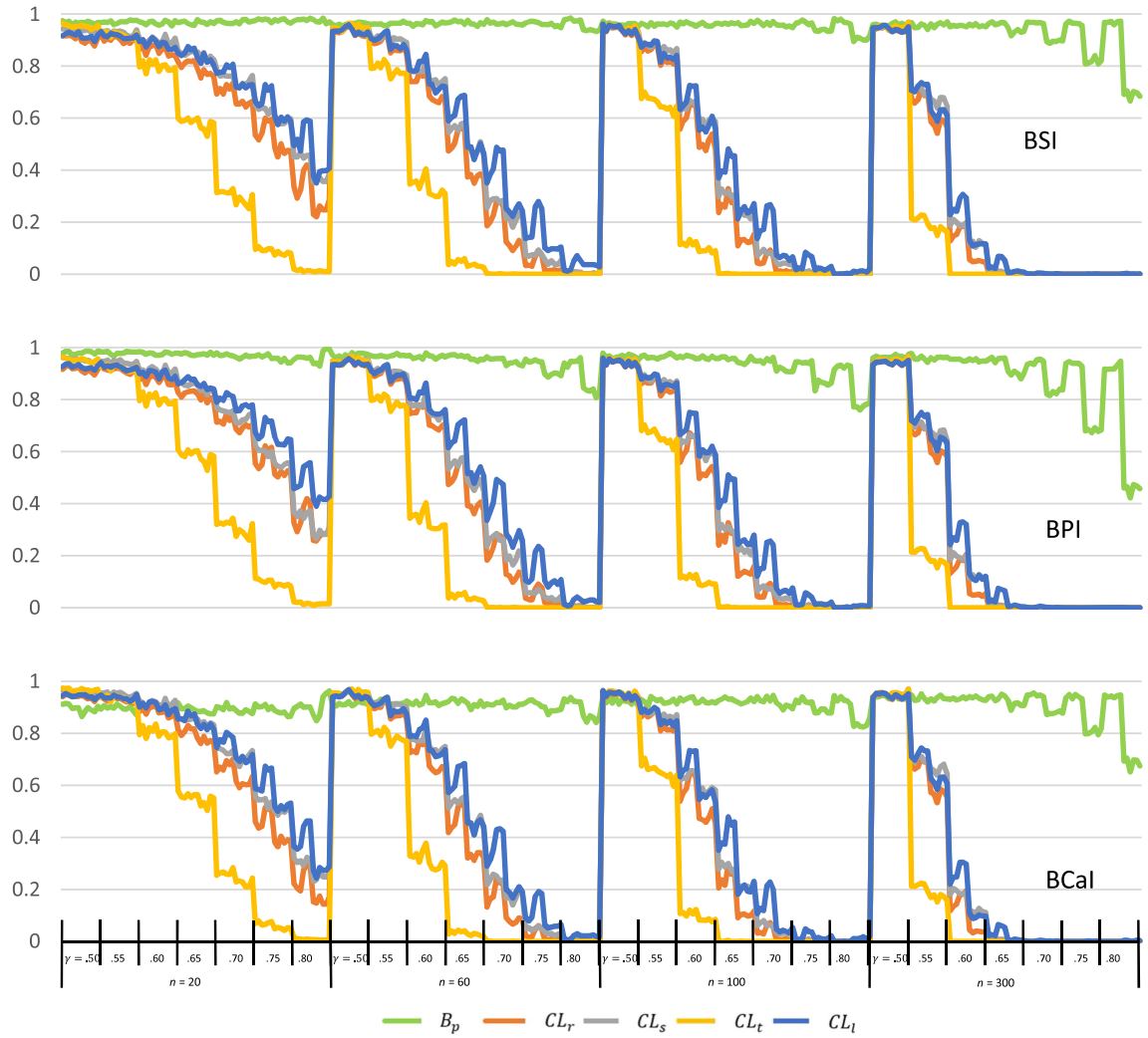

Fig. 6 Coverage probabilities when only the PBS condition is met (Type 2). BSI = bootstrap standard interval, BPI = bootstrap percentile interval, BCaI = bootstrap bias-corrected and accelerated percentile interval

BSI, BPI, and BCaI values are likewise less than optimal. For CL_r, the coverage probabilities yielded from BSI ranged from 0 to .957 with a mean of .443. Of the 336 conditions, only 29 (or 8.6%) fell within the criterion of [.925, .975]. For BPI, range = (0, .958), mean = .451, and 38 (11.3%) fell within the criterion. For BCaI, range = (0, .967), mean = .433, and 54 (or 16.1%) fell within the criterion. For CL_S, BSI produced a range of (0, .968), mean = .478, and 38 (11.3%) fell within the criterion; BPI yielded a range of (0, .968), mean = .477, and 64 (or 19.0%) fell within the criterion; BCaI resulted in a range of (0, .969), mean = .467, and 64 (or 19.0%) fell within the criterion. For CL_T, BSI produced a range of (0, .968), mean = .308, and 54 (16.1%) fell within the criterion; BPI yielded a range of (0, .966), mean = .309, and 54 (or 16.1%) fell within the criterion; BCaI resulted in a range of (0, .974), mean = .304, and 57 (or 17.0%) fell within the criterion. For CL_L, BSI produced a range of (0, .961), mean = .495, and 36 (10.7%) fell within the criterion; BPI yielded a range of (0, .959), mean = .503, and 48 (or 14.3%) fell within the criterion; BCaI resulted in a range of (0, .968), mean = .483, and 61 (or 18.2%) fell within the criterion. Because only the point estimates and BSI for B_P yielded reasonable results overall, the following discussion of the effects of the manipulated factors focuses only on the point estimates and BSI for B_P.

Effects of manipulated factors on B_P and BSI

The effects of the manipulated factors on the performance of B_p were minimal, as is shown in Table 2. There were no obvious effects of varying the distribution of Y on B_p and BSI. The most influential factor was the distribution of X (θ): When X is positively or negatively skewed, the point estimate biases were slightly more negative (except when $\gamma = .80$, $n = 20$, and θ = negatively skewed). For example, when $\gamma = .80$ and $n = 300$, the magnitude of the biases increased from – .009 (θ = normal) and – .008 (θ = uniform) to – .055 (θ = positively skewed) and – .054 (θ = negatively skewed). This may be due to the sample mean estimates ($\bar{x}$ and $\bar{y}$) in Eq. 7, which become less robust estimates of the center of the distribution when there is a long

tail. Second, when n was increased from 20 to 300, the accuracy of B_p improved, especially when the θ distribution is normal or uniform. This is reasonable, because a good sample estimate should be asymptotically assumed, meaning that when $n \rightarrow \infty$, $B_P \rightarrow \gamma$. Third, when γ increased from .50 to .80 and other factors were held constant, the biases of B_p became slightly more negative. This pattern is reasonable because γ has an upper bound of 1, which results in fewer sample B_p estimates above $\gamma = .80$ than below $\gamma = .80$. Generally, the effects of the manipulated factors on B_p are indeed minimal, and hence, these results demonstrate that B_P is a trustworthy estimator for γ across the conditions examined in this study.

For BSI, the two worst coverage probabilities (.687 and .695) resulted when $\gamma = .80$, $n = 300$, and θ = positively or negatively skewed. This is understandable because BSI depends upon the accuracy of the point estimate of B_p and under these conditions the least accurate point estimates were found. Also, these conditions are quite strict—that is, a relatively large $\gamma = .80$ (capped at 1), and a challenging skewed distribution (skewness = 4 or – 4 and kurtosis = 38), and a relatively narrow BSI (because of a large sample size). Hence, a narrow BSI becomes more sensitive to slight deviation from B_p to γ, and this inevitably results in smaller CPs. When other factors (n, γ, and θ = normal or uniform) were manipulated, the CPs of the BSI were robust across these conditions, demonstrating that BSI is a good CI construction procedure for B_p.

Working example

This section illustrates how researchers can obtain the B_p estimate of γ and its bootstrap CIs for their dataset using R (R Core Team, 2016; or RStudio: RStudio Team, 2016), a free and popular statistical package in behavioral and social sciences. We have made available the code for a function that computes the B_p estimate, together with a sample dataset and step-by-step instructions, in the supplementary materials. First, copy the code (function name: pbs) and execute it in R. Second, enter the X and Y scores from supplied in the supplementary materials to form a 100 × 2 data matrix in R. To best demonstrate how the code works, we have simulated the X and Y scores so that the population parameter γ is known and can be used to evaluate the accuracy of sample estimate B_P. In this example, the X and Y scores were generated from $\gamma = .60$, $n = 100$, θ has a uniform distribution, and $\sigma_Y = \sqrt{12}/2$. Third, run the syntax pbs(data, 1000, .95, 1234, 4) in R, where data refers to the name of the 100 × 2 data matrix, 1000 refers to the number of bootstrap samples, .95 is 95% CI, 1234 is the seed number, and 4 is the number of decimal places displayed in the output. If a researcher chooses to use these default settings, the syntax can further be simplified to: pbs(data). Alternatively, a researcher could alter any of the arguments to suit a study's particular needs (e.g., change the confidence interval to 99% by entering .99 in place of .95 in the third argument).

Once R finishes running the code, the results will be displayed (see Step 4 of the Appendix). In this case, $B_P = .61$, and we obtained a 95% BSI = (.5031, .7169), which indicates a statistically significant result at the .05 level because the range of the BSI confidence interval does not span .50. For purposes of interpretation, this $B_P = .61$ estimate tells us that there is a statistically significant 61% chance that when an X score is above the mean of all X scores, the paired Y score is also above the mean of all Y scores.

For purposes of comparison only (see note 2), computing the correlation for this simulated dataset produces r = .1210, p = .2306, which is nonsignificant at the .05 level. We compute the CL_r estimate of γ using Dunlap's conversion formula (Eq. 4) to get an estimate of .5386. The CL_r estimate is a biased estimate of γ, as should be expected, because the underlying X-Y relationship is not linear. Furthermore, a researcher that computes this estimate may mistakenly infer that because the correlation is not statistically significant there is also no significant PBS relationship between X and Y. However, using B_P to estimate γ and the 95% BSI to test for the statistical significance of this estimate, we can correctly identify a significant bivariate relationship that would be missed using traditional correlational analysis.

Real-world examples

For purposes of demonstration only,[3] this section presents how researchers could lead to different conclusions if they specify a different hypothesis (i.e., linearity vs. PBS) and use a different statistic (i.e., r vs. B_p) to test this hypothesis. We suggest that identifying PBS relationships in real world research can contribute valuable information not revealed by traditional correlational analysis. To demonstrate this contribution across different disciplines, we randomly selected two recently uploaded bivariate datasets for analysis from Ontario Data Documentation, Extraction Service and Infrastructure (ODESI), a Web-based digital repository for social sciences data.

The first dataset (Mychasiuk, 2017) contains behavioral data for 74 adolescent rats with mild traumatic brain injury (RmTBI)

[3] It is important to note that researchers should specify a core hypothesis prior to collecting data in any research studies. For example, if a researcher is interested in testing whether X and Y are linearly related, then the researcher should specify a hypothesis of a "linear relationship between X and Y" and use a corresponding statistic (e.g., r) to examine this hypothesis. If a researcher attempts to examine whether X and Y are PBS, then the researcher should test a hypothesis that "X and Y are PBS-related" and use a corresponding statistic (e.g., B_p) to verify this hypothesis. The present simultaneous examinations of an X–Y relationship through r and B_p in the working and real-world examples are done for the purposes of demonstration only, with the goal to explore and understand how the observed r and B_p values may differ in simulated and real-world datasets. In practice, researchers should keep in mind the importance of having a core or specific hypothesis before data collection and analysis. We thank an anonymous reviewer for pointing out this important point.

Table 2 Percentage biases of B_P and coverage probabilities of BSI in selected conditions

		$n = 20$		$n = 60$		$n = 100$		$n = 300$	
γ	θ	%Bias B_P	BSI	%Bias B_P	BSI	%Bias B_P	BSI	%Bias B_P	BSI
.50	1	.012	.972	.003	.959	.001	.967	.000	.959
	2	– .002	.964	– .001	.959	– .001	.961	.000	.956
	3	– .006	.969	– .003	.966	– .001	.959	.002	.965
	4	– .007	.965	.001	.965	– .002	.960	– .001	.962
.55	1	– .008	.968	– .004	.961	– .001	.959	– .001	.955
	2	– .006	.966	– .001	.959	– .002	.961	– .002	.954
	3	– .009	.971	– .011	.965	– .007	.961	– .007	.957
	4	– .017	**.976**	– .008	.968	– .009	.960	– .006	.965
.60	1	– .017	.966	– .007	.961	– .008	.968	– .003	.954
	2	– .014	.969	– .008	.966	– .006	.959	– .003	.954
	3	– .020	.965	– .018	.965	– .019	.962	– .015	.960
	4	– .025	.967	– .018	.965	– .016	.957	– .014	.953
.65	1	– .021	.970	– .010	.972	– .010	.964	– .006	.964
	2	– .020	.960	– .006	.968	– .010	.963	– .006	.965
	3	– .028	.972	– .026	.966	– .027	.960	– .024	.932
	4	– .031	.965	– .024	.960	– .025	.965	– .024	.935
.70	1	– .023	.973	– .013	.971	– .012	.966	– .008	.968
	2	– .024	.964	– .014	.963	– .012	.966	– .007	.966
	3	– .042	.971	– .035	.965	– .034	.958	– .033	**.893**
	4	– .040	.974	– .033	.960	– .035	.951	– .032	**.897**
.75	1	– .029	**.978**	– .018	.970	– .012	.970	– .007	.967
	2	– .025	.975	– .018	.966	– .014	.965	– .008	.962
	3	– .047	**.977**	– .044	.961	– .044	.937	– .043	**.812**
	4	– .048	.972	– .044	.952	– .042	.937	– .043	**.823**
.80	1	– .028	**.977**	– .019	**.982**	– .016	**.979**	– .009	.969
	2	– .026	**.980**	– .018	.975	– .015	.966	– .008	.967
	3	– .058	.973	– .054	.938	– .055	**.904**	– .055	**.687**
	4	.035	.972	– .056	.936	– .054	**.905**	– .054	**.695**

γ is the population PBS parameter, θ refers to the distribution type of X (1 = normal, 2 = uniform, 3 = positively skewed, and 4 = negatively skewed), n is the sample size, B_P is the sample estimator for γ, and BSI is the 95% bootstrap standard interval. Coverage probabilities outside acceptable range are presented in bold

after the consumption of caffeine. We obtained r and CL_r estimates for the relationship between weight (weightTBI) and the average time to right (AverageTimeToRight). A scatterplot of the data is shown in Fig. 7. The results show that $r = .158$, 95% CI = (– .083, .367), indicating no statistically significant linear relationship between these variables. Using Dunlap's (1994) transformation, we obtain $CL_r = .551$ as an estimate for γ. However, when directly computing an estimate for γ using our PBS method, we find $B_P = .581$, 95% BSI = (.455, .701). B_P estimates a stronger PBS relationship than does CL_r, although the 95% CI still spans the .50 PBS zero effect.

There is an important lesson we can take from this result. If we use Eq. 4 to convert our B_P estimate to the r-metric, the value is .252, which suggests a noticeably larger correlation than the actual r value obtained in the correlational analysis. This is not to suggest that the correlational analysis provided an underestimate of the linear relationship. To the contrary, we suggest that treating a measure of the PBS relationship as an effect size to describe a correlation is misleading and can result in precisely such incorrect inferences.

Dunlap's (1994) Eq. 4 implies that a PBS relationship requires the existence of a linear relationship and that the PBS estimate can describe the linear relationship. This is not the case. The present example demonstrates the type of problem that could arise as a result of treating CL_r or any other estimate of PBS as an effect size for a linear relationship. Although linearity implies the existence of a PBS relationship, the existence of a PBS relationship does not imply linearity. Being able to directly compute an estimate of the magnitude of a PBS relationship without dependence on

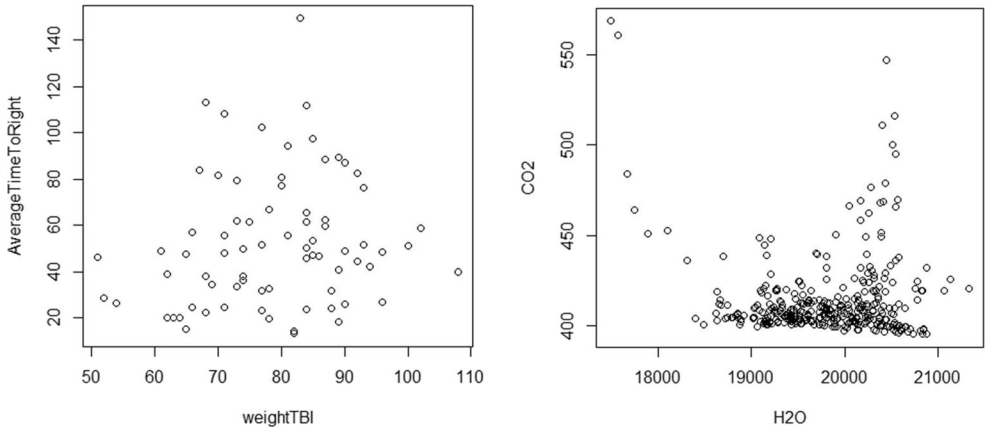

Fig. 7 Scatterplots for the real-world datasets. The left panel shows a scatterplot of the weight (weightTBI) of 74 adolescent rats to their average time to right (AverageTimeToRight) in the first real-world dataset (Mychasiuk, 2017). The right panel shows a scatterplot of concentrations of water (H_2O) and carbon dioxide (CO_2) in the second real-word dataset (Wunch et al., 2017)

correlational analysis should help reduce any confusion about the relationship between these bivariate forms. Our PBS algorithm makes reliance on the CL_r conversion formula unnecessary and will allow PBS to be appreciated independently of r.

The second dataset (Wunch, Arrowsmith, & Heerah, 2017) comes from an environmental study that measured the density of chemicals in a bike cargo trailer 362 times between June 28th to July 19th, 2017. We computed r and B_P estimates for the relationships between water (H_2O) and carbon dioxide (CO_2), $r = -.079$, 95% CI = (– .180, .024), which is a nonsignificant result. By conversion using Eq. 4, we found $CL_r = .475$, but we obtained a very different result by direct estimation, $B_P = .583$, 95% BSI = (.518, .648). A scatterplot of the data is shown in Fig. 7. Here we have identified a statistically significant PBS effect in the absence of a linear relationship. Relying upon correlational analysis, a researcher would conclude that there is no important relationship between H_2O and CO_2. The use of PBS analysis to obtain B_P leads to a very different inference: There is a significant probability-of-bivariate-superiority relationship between H_2O and CO_2, such that when the H_2O level is above the mean H_2O level, there is a 58.3% chance that the CO_2 level will be above the mean CO_2 level. This example illustrates the advantage of conducting PBS analysis when seeking to understand the relationship between two variables. Not only does PBS identify an important relationship when it exists, but it is also a relationship that is easy to communicate, making the dissemination of research findings more impactful.

Conclusions and discussion

In this article we have described a new statistical procedure to estimate the probability of bivariate superiority, an important type of bivariate relationship. Although little previous work has addressed this type of bivariate relationship, statisticians have suggested its importance under the framework of the copula theory (Jaworski et al., 2010). In addition, it has previously been suggested as a more understandable way to describe a bivariate relationship (Dunlap, 1994). PBS is not simply a concept translated from bivariate normal correlation (r), as is implied by Dunlap's CL_r. Rather, it is a unique theoretical and statistical model for quantifying bivariate relationships under the copula theory. PBS describes how likely it is that an X score that is above (or below) the mean is associated with a Y score that is above (or below) the mean.

In a Monte Carlo experiment we simulated data from bivariate relationships under a variety of conditions, many of which violated the parametric and linearity assumptions of conventional correlational analysis. We used our new PBS algorithm to compute a point estimate, B_P, of the true PBS, γ and confidence interval in each of 448 conditions. The results of the experiment demonstrate that B_P can appropriately identify this type of bivariate relationship, is robust to the simulation conditions that deviated from conventional parametric assumptions, and allows for inferences to be made as to the statistical significance of the bivariate relationship through the use of bootstrap CIs. Moreover, the likelihood-based interpretation of B_P is more understandable and interpretable than conventional r-based interpretations of bivariate relationships (i.e., proportion of variance explained).

The different B_P and CL_r estimates found in Distribution Type 1b suggest important implications about the use of CL_r in practice. These results suggest that CL_r estimates the PBS parameter γ differently than B_P, when the condition of linearity and nonnormality is met. Although we do not exactly know the true population γ value, we believe that the B_P procedure, which directly counts the number of times that $Y > \overline{Y}$ and $X > \overline{X}$ in a sample, appears to better measure and is more consistent with the concept of $\gamma = P\left(Y > \overline{Y} \cap X > \overline{X}\right)$ that defines the PBS parameter as compared to CL_r, which is based on the r-to-CL_r conversion, $\frac{1}{\pi} sin^{-1}(r) + 0.5$. In short, at least we are safe to

conclude that CL_r is of limited use when a researcher uses an r value and converts it to CL_r given that X and Y are linearly related but they are nonnormal. This result also supports Dunlap's (1994) suggestion that researchers can use the r-to-CL_r conversion, when X and Y follow bivariate normal correlation.

We provided a working example and an example from real world research that demonstrated that PBS can identify an important and significant bivariate relationship in the absence of a significant linear correlation. The correlation in the working example was nonsignificant at the .05 level, $r = .1210$, $p = .2306$. However, PBS analysis identified a bivariate superiority relationship, $B_P = .61$, $p < .05$. Critically, the examples we have outlined, together with our Monte Carlo results, imply that CL_r is not an adequate procedure for detecting PBS relationships because (a) a researcher that finds an r that is not statistically significant is unlikely to bother to transform that r to CL_r, (b) the transformation to CL_r provides no information about the significance of the effect unless an appropriate bootstrap technique is used to construct a confidence interval, and (c) when linearity and parametric assumptions are violated the transformation from r to CL_r leads to biased estimates of γ that could lead to erroneous inferences. Hence, researchers that have relied upon r or its related models (e.g., linear regression model) to evaluate bivariate relationships, even those that have transformed their results to CL_r, may have missed PBS relationships that are important for theory testing and model building in behavioral and social sciences.

We have outlined a solution to these problems and have specified a reliable method to directly compute a point estimate of γ that is robust to violations of parametric assumptions and provides results that are more easily communicated for research dissemination. We have proposed PBS as a new statistical tool that can be used in future research to identify the probability of superiority in bivariate relationships. In addition, the effect size estimate produced by PBS analysis, B_P, is a common-language effect size that can make communicating the character of the bivariate relationship more successful. Finally, we propose PBS as a statistical analysis that behavioral and social science researchers can apply to past research. We encourage researchers to reexamine bivariate relationships in their datasets to find theoretically important effects that had been previously overlooked.

Limitations and future directions

We were inspired by the ground-breaking work of others (Dunlap, 1994; Grissom, 1994; McGraw & Wong, 1992; Vargha & Delaney, 2000; Wolfe & Hogg, 1971) to develop PBS, and we constructed this new procedure upon the solid foundations they built. However, PBS is a new statistical procedure, and there is much work to be done to provide a more complete picture of PBS and its theoretical and practical applications. The nature of conducting a Monte Carlo experiment is that one must choose from among many variables that can be manipulated, and choose the levels at which each variable will be tested. Although we are confident that we have chosen the most important variables for an initial test of our procedure, and levels for each that are adequately representative of many common data circumstances encountered in real research scenarios, we recognize that there are other conditions of import under which PBS should be evaluated. In particular, additional nonparametric distributions of X (e.g., bimodal, U-quadratic, normal-ogive, logistic) and how this "glues" to γ with a particular level of γ should be considered in future research.

A priority for advancing PBS, both theoretically and practically, is the development of generalized forms of B_P for use in research scenarios involving more than two variables. Behavioral and social science researchers often investigate how an outcome measure (or criterion) can be regressed on multiple predictor variables through regression analysis. Additional research is necessary to examine how the PBS concept can be extended and generalized to more complex research situations of this nature.

Another future direction for the development of PBS involves examination of the diagnostic value of the probability-of-superiority conceptualization. The basic idea of PBS focuses on the likelihood that when an X score (e.g., daily exercise) is above (or below) the mean, the paired Y score (e.g., hospitalization) is also above (or below) the mean. Consequently, researchers may use PBS information to classify or diagnose individual participants into a 2 × 2 profile—that is, daily exercise (good or bad) and chance of hospitalization (high or low). Future research is necessary to investigate the accuracy and usefulness of these types of PBS-implied diagnostic profiles for each individual participant in a research study. We expect this approach will lead to development of a PBS-based method for individualized diagnostic information to be communicated to people in a way that is both understandable and useful to them.

Whereas simulation studies, such as the Monte Carlo experiment we have presented herein, are effective for demonstrating the performance of a statistical procedure under a wide range of conditions, they are often not sufficient to convince the cautious research community to adopt new techniques of analysis. Therefore, it is necessary to begin applying PBS to existing datasets in the behavioral and social sciences literature. We invite researchers to undertake this task independently, and we invite collaborations, to explore the degree to which significant PBS relationships have been missed in previously published and unpublished datasets. This undertaking will accomplish three important objectives: (a) It will

provide a proving ground for PBS in real-data scenarios, (b) it will enable researchers to familiarize themselves with the PBS procedure and interpretive structure, and (c) it will allow researchers to identify previously overlooked bivariate relationships of theoretical importance to their research.

Author note This research was supported by the University Research Grants Program (URGP) to Johnson Ching-Hong Li in the Department of Psychology at the University of Manitoba (#47094).

Appendix

Following the mathematical proof from Blomqvist (1950), we can derive a mathematical relationship between CL_r and B_p, when the condition of bivariate normal correlation is met.

Assume that (x_i, y_i), where $i = 1, 2, \ldots, n$, are n samples from a two-dimensional population associated with a bivariate normal, correlational cdf $f(x, y)$,

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

where x is an observation in variable X, y is an observation in variable Y, r is the sample correlation coefficient, $\bar{x}$ is the sample mean of x, $\bar{y}$ is the sample mean of y, s_x is the sample SD of X, s_y is the sample SD of Y, and $\pi \approx 3.14156$. Let the x-plane and y-plane be divided into four quadrants based on the lines $x = \bar{x}$ and $y = \bar{y}$. The cdf $f(x, y)$ is assumed to have continuous marginal $f(x)$ and $f(y)$ such that the probability of obtaining two equal x-values or two equal y-values in a sample will be zero. Consequently, this implies that some information about correlation coefficient r can be obtained and mathematically linked to a probability-based estimate (called q in Blomqvist, 1950) based on 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.

When the bivariate normal correlation is met, Blomqvist's (1950) Eq. 12 provides a mathematical proof between q and r,

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

where "≡" refers to the equal sign given the definition of the bivariate normal correlation, and $\pi \approx 3.14156$. Dividing Eq. 1 by 2 and adding a constant 0.5, Eq. 1 becomes Eq. 2,

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

According to Blomqvist's equation (Eq. 1), there exists an estimator (called q') that can estimate the aforementioned probability-based parameter q, without the condition of bivariate normal correlation. That is, q is estimated by q', which is equal to

$$q' = \frac{n_1 - n_2}{n_1 + n_2} = \frac{2n_1}{n_1 + n_2} - 1, \tag{3}$$

where n_1 refers to the number of sample data points (x, y) belong to the first or third quadrants, and n_2 refers to the number of sample data points (x, y) belong to the second or fourth quadrants.

Extracting the algorithm from the left-hand side of Eq. 1 and substitute it by Eq. 3, we obtain

$$\begin{aligned} q = q' &= \frac{n_1 - n_2}{n_1 + n_2} = 2 \cdot \frac{n_1}{n_1 + n_2} - 1 \\ &= 2 \cdot P(Y > \bar{Y} \cap X > \bar{X}) - 1, \end{aligned} \tag{4}$$

Solving for $P\left(Y > \overline{Y} \,\middle|\, X > \overline{X}\right)$ in Eq. 4, we obtain

$$\frac{n_1}{n_1 + n_2} = P\left(Y > \overline{Y} \cap X > \overline{X}\right) = q\left(\frac{1}{2}\right) + 0.5, \tag{5}$$

where $P\left(Y > \overline{Y} \cap X > \overline{X}\right) = \frac{n_1}{n_1 + n_2}$ refers to the probability-of-bivariate-superiority (PBS) estimate.

For continuous marginal $f(x)$ and $f(y)$, and when the probability of obtaining two equal x-values or two equal y-values in a sample is zero, our proposed B_P can be linked to $P\left(Y > \overline{Y} \cap X > \overline{X}\right)$,

$$\begin{aligned} B_P &= \frac{\sum_{i=1}^{n} \#\left[\text{sign}\left(x_i - \bar{x}\right) \cdot \text{sign}\left(y_i - \bar{y}\right) > 0\right] + 0.5 \#\left[\text{sign}\left(x_i - \bar{x}\right) \cdot \text{sign}\left(y_i - \bar{y}\right) = 0\right]}{n_1 + n_2} \\ &= \frac{\sum_{i=1}^{n} \#\left[\text{sign}\left(x_i - \bar{x}\right) \cdot \text{sign}\left(y_i - \bar{y}\right) > 0\right]}{n_1 + n_2} \\ &= \frac{n_1}{n_1 + n_2} = P\left(Y > \overline{Y} \cap X > \overline{X}\right) = q\left(\frac{1}{2}\right) + 0.5. \end{aligned} \tag{6}$$

It is important to note that the proof from Eqs. 3 to 6 does not depend upon the condition of bivariate normal

correlation, and hence, our proposed B_P is considered a robust estimator for $P(Y > \overline{Y} \cap X > \overline{X})$.

When we attempt to provide a mathematical relationship between B_P and CL_r, the condition of bivariate normal correlation is necessary for developing such a relationship. According to Eq. 6,

$$B_P = q\left(\frac{1}{2}\right) + 0.5. \tag{7}$$

Substitute Eq. 7 into Eq. 2, with the condition of bivariate normal correlation, we derive that B_P can be mathematically linked to Dunlap's (1994) CL_r (and correlation coefficient r) by

$$\begin{aligned} B_P &= q\left(\frac{1}{2}\right) + 0.5 \\ &= \frac{\sum_{i=1}^{n} \#\left[\text{sign}\left(x_i - \overline{x}\right)\cdot\text{sign}\left(y_i - \overline{y}\right) > 0\right] + 0.5\#\left[\text{sign}\left(x_i - \overline{x}\right)\cdot\text{sign}\left(y_i - \overline{y}\right) = 0\right]}{n} \\ &\equiv \frac{1}{\pi} sin^{-1}(r) + 0.5 = CL_R, \end{aligned} \tag{8}$$

if and only if the condition of bivariate normal correlation is met.

References

Blomqvist, N. (1950). On a measure of dependence between two random variables. *Annals of Mathematical Statistics*, *21*, 593–600.

Botev, Z. I. (2017). The normal law under linear restrictions: Simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *79*, 125–148. https://doi.org/10.1111/rssb.12162

Bradley, J. (1982). The insidious L-shaped distribution. *Bulletin of the Psychonomic Society*, *20*, 85–88.

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

Canty, A., & Ripley, B. (2016). boot: Bootstrap R (S-Plus) functions (R package version 1.3-18). Retrieved from https://cran.r-project.org/web/packages/boot/index.html

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*, 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*, 494–509. https://doi.org/10.1037/0033-2909.114.3.494

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

Dunlap, W. P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin*, *116*, 509–511. https://doi.org/10.1037/0033-2909.116.3.50

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

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*, 1–51. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

Grissom, R. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, *79*, 314–316.

Hogg, R., & Craig, A. (1971). *Introduction to mathematical statistics* (4th ed.). New York, NY: Macmillan.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Wadsworth.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement*, *60*, 543–563. https://doi.org/10.1177/0013164400604004

Jaworski, P., Durante, F., Härdle, W. K., & Rychlik, T. (Eds.). (2010). *Copula theory and its applications: Proceedings of the workshop held in Warsaw, 25–26 September 2009*. Berlin, Germany: Springer.

Karl, P. (1895). VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society*, *58*, 240–242. https://doi.org/10.1098/rspl.1895.0041

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, *30*, 81–89. https://doi.org/10.1093/biomet/30.1-2.81

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed.). New York, NY: Macmillan.

Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). London, UK: Edward Arnold

Lai, C., & Balakrishnan, N. (2009). *Continuous bivariate distributions*. New York, NY: Springer.

Leech, N. L., & Onwuegbuzie, A. J. (2002). *A call for greater use of nonparametric statistics*. Retrieved from files.eric.ed.gov.login.ezproxy.library.ualberta.ca/fulltext /ED471346.pdf

Li, J. C.-H. (2015). 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., Chan, W., & Cui, Y. (2011). Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. *British Journal of Mathematical and Statistical Psychology*, *64*, 367–387. https://doi.org/10.1348/2044-8317.002007

Ling, Y., & Nelson, P. I. (2014). Effect sizes for comparing two or more normal distributions based on maximal contrasts in outcomes. *Statistical Methods & Applications*, *23*, 381–399. https://doi.org/10.1007/s10260-014-0254-y

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

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Mychasiuk, R. (2017). *Behavioral and pathophysiological outcomes associated with caffeine consumption and repetitive mild traumatic brain injury (RmTBI) in adolescent rats* (Scholars Portal Dataverse, V1). doi:10.5683/SP/8RODEV

Nelson, R. B. (2006). *An introduction to copulas* (2nd ed.). New York, NY: Springer.

Onwuegbuzie, A. J., & Daniel, L. G. (2002). Uses and misuses of the correlation coefficient. *Research in the Schools*, *9*, 73–90.

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

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*, 1518–1524. https://doi.org/10.1126/science.1205438

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician*, *42*, 59–66. Retrieved from www.jstor.org/stable/2685263

Royal Statistical Society. (2010). *Statistical literacy*. Retrieved from www.rss.org.uk/RSS/Influencing_Change/Statistical_literacy/RSS/Influencing_

Change/Statistical_literacy.aspx?hkey=821bf2f4-8a09-413c-8d22-290e2209a92a

RStudio Team. (2016). RStudio: Integrated development for R (website). Boston, MA: RStudio, Inc. Retrieved from www.rstudio.com

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

Siegal, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York, NY: McGraw-Hill.

Tomitaka, S., Kawasaki, Y., Ide, K., Yamada, H., Miyake, H., & Furukaw, T. A. (2016). Distribution of total depressive symptoms scores and each depressive symptom item in a sample of Japanese employees. *PLoS ONE*, *11*, e0147577. https://doi.org/10.1371/journal.pone.0147577

United Nations Economic Commission for Europe. (2009). *Making data meaningful*. Retrieved from https://www.unece.org/fileadmin/DAM/stats/documents/writing/ Making_Data_Meaningful_Part_4_for_Web.pdf

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*, 101–132.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Amsterdam, The Netherlands: Elsevier.

Wilcox, R. R., Granger, D. A., Szanton, S., & Clark, F. (2014). Cortisol diurnal patterns, associations with depressive symptoms, and the impact of intervention in older adults: Results using modern robust methods aimed at dealing with low power due to violations of standard assumptions. *Hormones and Behavior*, *65*, 219–225.

Wolfe, D. A., & Hogg, R. V. (1971). On constructing statistics and reporting data. *American Statistician*, *25*, 27–30.

Wunch, D., Arrowsmith, C., & Heerah, S. (2017). *GTA bike surveys June 28–July 19, 2017* (Scholars Portal Dataverse, V1). https://doi.org/10.5683/SP/ZDK98D