Behav Res (2016) 48:1560–1574
DOI 10.3758/s13428-015-0667-z

Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data

Johnson Ching-Hong Li[1]

Published online: 20 October 2015

Abstract In psychological science, the "new statistics" refer to the new statistical practices that focus on effect size (ES) evaluation instead of conventional null-hypothesis significance testing (Cumming, *Psychological Science*, *25*, 7–29, 2014). In a two-independent-samples scenario, Cohen's (1988) standardized mean difference (*d*) is the most popular ES, but its accuracy relies on two assumptions: normality and homogeneity of variances. Five other ESs—the unscaled robust *d* (d_r^*; Hogarty & Kromrey, 2001), scaled robust *d* (d_r; Algina, Keselman, & Penfield, *Psychological Methods*, *10*, 317–328, 2005), point-biserial correlation (r_{pb}; McGrath & Meyer, *Psychological Methods*, *11*, 386–401, 2006), common-language ES (*CL*; Cliff, *Psychological Bulletin*, *114*, 494–509, 1993), and nonparametric estimator for *CL* (A_w; Ruscio, *Psychological Methods*, *13*, 19–30, 2008)—may be robust to violations of these assumptions, but no study has systematically evaluated their performance. Thus, in this simulation study the performance of these six ESs was examined across five factors: data distribution, sample, base rate, variance ratio, and sample size. The results showed that A_w and d_r were generally robust to these violations, and A_w slightly outperformed d_r. Implications for the use of A_w and d_r in real-world research are discussed.

Keywords Common-language effect size · Nonnormal data · Unequal variances · Simulation

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

[1] Department of Psychology, University of Manitoba, P517B, Duff Roblin Building, Winnipeg, Manitoba R3T 2N2, Canada

The "new statistics," an innovative framework developed by a number of methodological and quantitative researchers (as is detailed by Cumming, 2014), refers to new recommended practices that arose in response to perceived flaws in conventional, widely employed null-hypothesis significance testing (NHST). In NHST, when a researcher examines whether or not a significant mean difference exists in a dependent variable (DV; e.g., communication skills) between two groups of participants (e.g., female and male groups), the researcher often uses independent-samples *t* tests to obtain an observed probability (*p*) value. When an observed *p* is less than .05 (i.e., $p < .05$), the chance for observing such a large mean difference (e.g., female and male employees differ in cognitive skills by one standard deviation) is very unlikely, if their underlying true means are the same in the population. On the basis of this result, the researcher concludes that the observed mean difference is statistically significant at the .05 level because the difference is very likely to have arisen from different underlying populations.

This strategy, however, has perceived flaws. One can obtain a significant result with a large sample size even with a very small effect size (ES), which is a quantity that directly measures the strength of the association or difference between variables. Thus, the new statistical practices shift from dependence on NHST to reporting of an ES and its confidence interval (CI). Researchers thus can directly report the strength or magnitude of a relationship and its sampling error, without worrying about the impact of a large sample size on NHST. In addition, reporting the CI makes reporting the significance level redundant. In fact, many methodologists and journal editors have suggested that researchers should report their ES in order to quantify the strength of a relationship and should provide the associated CI in order to present the range of possible ESs that are likely to be obtained (i.e., sampling error) if a similar study were replicated in the future (e.g., Fritz, Morris, & Richler, 2012; Kline, 2013). The American Psychological Association

(APA) also strongly recommends reporting of the ES and CI: "estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all APA journals" (APA, 2010, p. 33). In addition, ES is an important statistic in meta-analysis, a popular statistical method that involves pooling the ESs to summarize the overall magnitude across studies conducted by independent researchers (Schmidt & Hunter, 2014).

There are a number of ES measures for the two-independent-samples case with one grouping variable (e.g., gender) and one continuous variable (e.g., communication skills). One of the most popular ES measures is Cohen's *d* (Cohen, 1988), which measures the separation (mean difference) between two groups or samples of observations, divided by the pooled standard deviation (*SD*). In an equation,

$$d = \left(\overline{Y}_1 - \overline{Y}_2\right)/s_p, \quad (1)$$

where $\overline{Y}_1$ and $\overline{Y}_2$ are the mean scores in Groups 1 and 2, respectively, and s_p is the pooled *SD* of Groups 1 and 2—that is, $s_p = \sqrt{\left[(n_1-1)s_1^2 + (n_2-1)s_2^2\right]/(n_i + n_2-2)}$, where n_i and s_i are the sample size and *SD* of observations in group $i = 1, 2$, respectively. If the female and male employees differed in their communication skills by one standardized unit in a sample, one could report $d = 1.00$ to express the magnitude of difference in communication skills between these two groups. According to Cohen, the interpretation for a small, moderate, and large ES is $d = 0.20$, 0.50, and 0.80. Note that *d* is unaffected by a large sample size when other factors are held constant, and hence, *d* adheres to the new statistical practices and is widely accepted among researchers.

Despite the popularity of *d*, its accuracy relies on two key assumptions about the underlying populations—normality and the homogeneity of variances—that may be violated in practice. *Normality* means that measurements of the DV in the underlying population are normally distributed. The assumption of homogeneity of variances is based on the notion that the variances of the DV should be the same in the two groups. Data in the behavioral and social sciences, however, often deviate from these assumptions. This may lead to inaccurate interpretation of ES, which, in turn, hinders the progress of the new statistical practices when they rely on *d*. Five other ES measures in the literature may be insensitive or robust to violations of these assumptions: the unscaled robust *d* (d_r^*; Hogarty & Kromrey, 2001), scaled robust *d* (d_r; Algina, Keselman, & Penfield, 2005), point-biserial correlation (r_{pb}; McGrath & Meyer, 2006), common-language ES (*CL*; Cliff, 1993), and nonparametric estimator for *CL* (A_w; Ruscio, 2008). However, no study has systematically and comprehensively examined the performance of these ES measures in one simulation study. Thus, little guidance is available to help researchers determine the most accurate ES to report and interpret under different data conditions.

The purpose of this study is to fill in this research gap by evaluating the performance of the six ES measures on the basis of a Monte Carlo simulation study, a widely used strategy for examining the overall performance of a statistical method across simulated and replicated samples in a computerized statistical package. The objectives of this study are (1) to systematically evaluate the accuracy of the six ES measures and (2) to provide recommendations for reporting and interpreting the most appropriate ES under different data conditions.

This article is divided in five sections. The first section discusses the assumptions for *d*. The second section presents the defining and computational details of other ESs that appear to be insensitive or robust to violations of data assumptions. The third section explains the methods and design of the Monte Carlo study. In the fourth section, the performance of the ESs is explained and evaluated on the basis of the simulation results. The fifth section discusses the implications of these ESs to real-world applications.

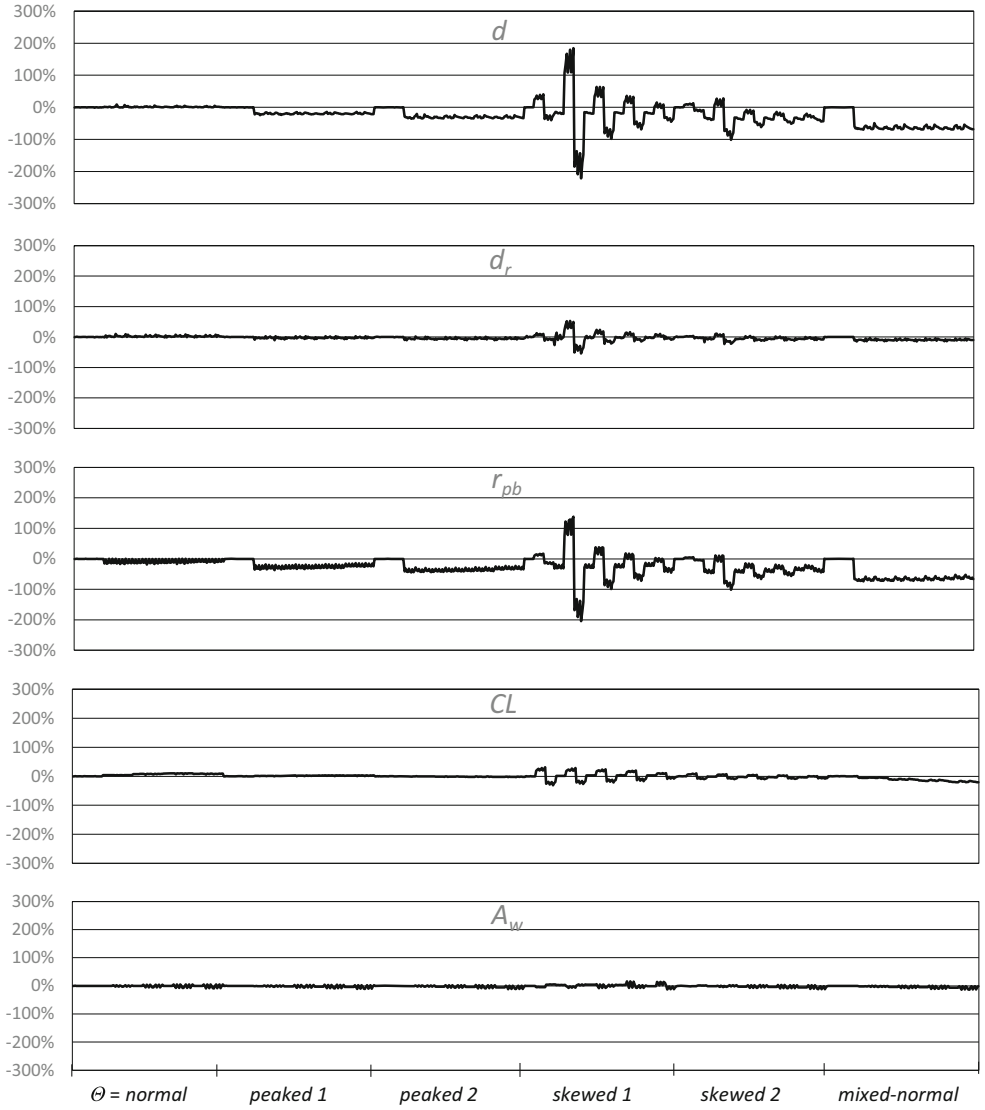
Data assumptions for *d*

Normality

Normality means that measures of the DV are independently and normally distributed in the underlying population. Data in behavioral science, however, often deviate from this assumption. For instance, data observed in some populations (e.g., clinical patients, gifted children) tend to follow a heavy-tailed (i.e., skewed) distribution. According to Algina et al. (2005), a mixed-normal distribution is also common in behavioral science. That is, a proportion (e.g., 10 %) of observations may come from a different normal distribution [e.g., *N*(0, 10); i.e., a normal distribution with a mean of 0 and an *SD* of 10] instead of the conventional *N*(0, 1), meaning that the shape of the mixed-normal distribution looks like the standard normal distribution, but it has a longer tail on both ends of the bell-shaped curve. The mixed-normal distribution can be found in a sample (e.g., a big school with lots of students) that has a mixture of very high and very low scorers (e.g., low and high achievers, giving both positive and negative outlier scores). The visual similarity between the normal and mixed-normal distributions often creates the illusion that the data have met the normal condition, which researchers are usually not aware of (see Fig. 1 in the Results; Algina et al., 2005). Unfortunately, Algina et al. found that *d* is not robust to a mixed-normal distribution, causing serious flaws in interpreting the ES.

Homogeneity of variances

Homogeneity of variances requires that the variances of the observations not be different for the two groups. Algina et al. (2005) found that the observed *d* becomes inaccurate when the



Note: d is Cohen's d, d_r is the scaled robust d, d_r^* is the non-scaled robust d, r_{pb} is point-biserial correlation, CL is the common language effect size, and A_w is the non-parametric estimator for CL. Θ corresponds to the normal ($\Upsilon_1 = 0; \Upsilon_2 = 0$), two peaked ($\Upsilon_1 = 0; \Upsilon_2 = 6$ and $\Upsilon_1 = 0; \Upsilon_2 = 154.84$), two skewed ($\Upsilon_1 = 2; \Upsilon_2 = 6$ and $\Upsilon_1 = 4.90; \Upsilon_2 = 4{,}673.80$), and, mixed-normal distributions ($\Upsilon_1 = 0; \Upsilon_2 = 24.95$).

Fig. 1 Percentage biases of the effect size measures across 810 simulation conditions. In the figure, d is Cohen's d, d_r is the scaled robust d, d_r^* is the nonscaled robust d, r_{pb} is point-biserial correlation, CL is the common-language effect size, and A_w is the nonparametric estimator for CL. Θ corresponds to the normal distribution ($\Upsilon_1 = 0$; $\Upsilon_2 = 0$), two peaked distributions ($\Upsilon_1 = 0$; $\Upsilon_2 = 6$ and $\Upsilon_1 = 0$; $\Upsilon_2 = 154.84$), two skewed distributions ($\Upsilon_1 = 2$; $\Upsilon_2 = 6$ and $\Upsilon_1 = 4.90$; $\Upsilon_2 = 4{,}673.80$), and a mixed-normal distribution ($\Upsilon_1 = 0$; $\Upsilon_2 = 24.95$)

variance ratio becomes 1:4 between the two groups. Moreover, heterogeneity should imply different conceptions of ES, because each group produces a distribution with different variance and shape. Cohen's d, a measure of location of separation between two groups of observations, cannot precisely reflect the differences in the scores between the two differently shaped distributions. In behavioral research, however, violations of the homogeneity of variances are not uncommon. Wilcox (1987) found that ratios of the largest to the smallest sample variances (i.e., variance ratios; VRs) that exceed 16 are not uncommon. In clinical research, significant differences in variances are usually found between treatment and control groups (Weisz, Weiss, Han, Granger, & Morton, 1995). In a published issue of the *Journal of Consulting and Clinical Psychology* (Brown, Evans, Miller, Burgess, & Mueller, 1997), Grissom and Kim (2001) found that VRs could range from 3.24 to 284.79 when they compared the variance of behavior-avoidance scores between a systematic desensitization group and a control group. In another study, Ruscio and Roche (2012) recorded the within-groups variances of the DVs

reported in 455 studies published in top-tier journals in psychology (e.g., *Journal of Applied Psychology*, *Journal of Educational Psychology*). The authors found that the majority of the sample variances differed substantially between groups of participants, thereby implying that the homogeneity-of-variance assumption is frequently violated in practice.

Other ES measures

Robust ds (d_r^* and d_r)

Hogarty and Kromrey (2001) and Algina et al. (2005) proposed and developed d_r^* and d_r, respectively, within the theory of robust statistics. Robust statistical methods often involve removing a proportion (e.g., w = 20 %; for an explanation of the value 20 %, see Wilcox, 2005) of high and low scores in a sample. This, in turn, eliminates the problem of outliers in each group that usually leads to extreme variances and skewed distributions. The first ES, d_r^*, is the unscaled robust estimator for d (Hogarty & Kromrey, 2001)—that is,

$$d_r^* = \left(\overline{Y}_{t1} - \overline{Y}_{t2}\right)/s_{tp}, \tag{2}$$

where $\overline{Y}_{t1}$ and $\overline{Y}_{t2}$ are the 20 % trimmed means for Groups 1 and 2, respectively, and s_{tp} is the square root of the pooled 20 % Winsorized variance—that is, $s_{tp} = \sqrt{\left[(n_1-1)s_{t1}^2 + (n_2-1)s_{t2}^2\right]/(n_1+n_2-2)}$, where n_i and s_{ti}^2 are the sample size and 20 % Winsorized variance[1] of the observations in group i = 1, 2, respectively. The second ES, d_r, is the scaled robust estimator for d (Algina et al., 2005)—that is,

$$d_r = .642 \cdot d_r^*. \tag{3}$$

Algina et al. stated that it is not necessary to multiply d_r^* by .642 to produce d_r, although such a scale-multiplication could compensate for the impact of removing a proportion of the observations and transform d_r^* to d_r, such that d_r is measured on the same standardized mean difference metric as d, which is common in many robust statistical methods. Algina et al. found that the coverage probabilities yielded by the bootstrap CIs surrounding d_r were accurate. On the other hand, Hogarty and Kromrey (2001) investigated the accuracy of d_r^* and found that the results were generally reasonable.

[1] The Winsorized variance is the variance of observations, in which a w% (e.g., 20 %) of the top and bottom scores are dropped and replaced by the highest and lowest scores, respectively, in the remaining sample. Consider a sample that contains {1, 1, 2, 3, 3, 4, 4, 5, 6, 7}; its Winsorized variance becomes the variance of the Winsorized sample—that is, {2, 2, 2, 3, 3, 4, 4, 5, 5, 5}.

Point-biserial correlation (r_{pb})

Another conventional ES for the two-independent-samples case is r_{pb}, which is mathematically equivalent to Pearson's correlation when applied to one grouping variable and one numeric variable. In an equation,

$$r_{pb} = \sqrt{pq} \cdot \left(\overline{Y}_1 - \overline{Y}_2\right)/s_Y, \tag{4}$$

where p and q are the proportions of observations in Groups 1 and 2, respectively, $\overline{Y}_1$ and $\overline{Y}_2$ are the means of Groups 1 and 2, respectively, and s_Y is the *SD* of all observations in *Y*. Because r_{pb} is a derivative of r, the usual assumptions, normality and continuality, are required. In addition, r_{pb} is sensitive to the ratio of the sample sizes between two groups (i.e., base rate), as is evidenced by the term $\left(\sqrt{pq}\right)$ in Eq. 4. It is, however, unknown whether or not homogeneity of variances is necessary for r_{pb}, because s_Y measures the variability of all Y scores regardless of their group memberships. McGrath and Meyer (2006) compared the differences between r and d and offered recommendations for researchers to choose which ES to report. r_{pb} is particularly useful in cases in which the goal is to evaluate criterion-related validity, whereas d is more suited to scenarios in which the goal is to evaluate the effect of an experiment or intervention. Note that r_{pb} is mathematically related to d—that is,

$$r_{pb} = d/\sqrt{d^2 + (1/pq)}. \tag{5}$$

Parametric common-language ES (CL)

Cliff (1993) was one of the pioneer studies that proposed the use of the common-language ES measure (*CL*), which aimed to communicate an ES measure in a manner understandable by laypersons. *CL* makes use of the parameter $Pr(Y_1 > Y_2)$, which measures the probability that a randomly selected score in Group 1 is higher than a randomly selected score in Group 2. For example, when a researcher compares the difference in subjective well-being (SWB) between a treatment group and control group, the *CL* estimates the probability that someone who receives the treatment would have greater SWB than someone in the control group. When the data meet the assumptions of normality within groups, *CL* can be estimated by

$$CL = \Phi\left[\left(\overline{Y}_1 - \overline{Y}_2\right)/s_p\right], \tag{6}$$

where Φ is the normal cumulative distribution function, $\overline{Y}_i$ is the mean of observations in group i = 1, 2, respectively, and s_p is the pooled *SD* as defined in Eq. 1. When the normality assumption is met and the samples sizes are equal, the criteria

for a small, a moderate, and a large ES for *CL* are 0.56, 0.64, and 0.71, respectively, which corresponds to 0.20, 0.50, and 0.80 in *d* (see note 2).

Nonparametric estimator for CL (A_w)

The probability of superiority ES measure (A_w), a nonparametric complement to the parametric *CL*, has received increasing attention in behavioral science (e.g., Delaney & Vargha, 2002; Grissom, 1994; Grissom & Kim, 2001, 2005; Hsu, 2004; McGrath & Meyer, 2006; Ruscio, 2008; Vargha & Delaney, 2000). Theoretically, A_w does not require the assumptions of normality and homogeneity of variances, but its robustness to these violations needs further empirical testing. In an equation, A_w expresses ES on the basis of the probability that a random observation of population *p* scores higher than a random observation of population *q*—that is,

$$A_w = [\#(\boldsymbol{p} > \boldsymbol{q}) + .5\#(\boldsymbol{p} = \boldsymbol{q})]/n_p n_q, \tag{7}$$

where # is the count function, $\boldsymbol{p}$ and $\boldsymbol{q}$ are vectors of scores for the two samples, and n_i is the sample size in group $i = p, q$. Consider $\boldsymbol{p}$ = {5, 7, 6, 5} and $\boldsymbol{q}$ = {3, 4, 5, 3}, the count function—#($p = 5 > q = 3, 4, 5, 3$)—yields a total count of 3.5. Repeat this process for the remaining elements in $\boldsymbol{p}$, A = (3.5 + 4 + 4 + 3.5)/16 = .9375, meaning that there is a 93.75 % chance that the observation would be higher for a randomly selected member of group *p* than for a randomly selection member of group *q*. Ruscio (2008) found that the nonparametric A_w was generally accurate and suggested that researchers and practitioners should report this measure; however, the question of its improved accuracy relative to the other five ESs needs further examination.

In light of the six different ESs available for a two-independent-samples case, it is crucial for researchers and practitioners to report and interpret the most appropriate ES, especially when their data violate the assumptions of normality and homogeneity of variances, a situation common in behavioral science research. On one hand, the robust d_r^*, d_r, and A_w appear to be robust to violations of normality and homogeneity. On the other hand, robust statistics usually require more observations than the conventional, parametric statistics (e.g., d, r_{pb}) to maintain the same level of accuracy as when the assumptions are met. Hence, a simulation study is required to examine the pros and cons of reporting different ESs across different data conditions. Therefore, the purpose of this study is to fill in the gap by examining the performance of the six ES measures in a Monte Carlo simulation study. This study was designed to systematically evaluate the performance of the six ES measures and offer recommendations to researchers and practitioners for the reporting and interpretation of the most appropriate ES under different data conditions.

Although the present study focuses on examining the accuracy of the point estimates of the six ESs, many journal editorials or publication manuals (e.g., APA, 2010) strongly recommend the reporting of both an ES and its CI. Hence, a description of how to construct the CIs for the six ESs is provided in the Appendix to serve as a practical guideline for researchers.

Method

A Monte Carlo study was conducted to systematically evaluate the performance of d, d_r^*, d_r, r_{pb}, *CL*, and A_w under the following simulated conditions.

Factor 1: Distribution (Θ; six levels). The first distribution follows a normal distribution [$N(1, 0)$] with skewness (Υ_1) = 0 and kurtosis (Υ_2) = 0. The following nonnormal (i.e., peaked and skewed) distributions were generated on the basis of Algina et al. (2005), in which the generated normal data were multiplied by particular g and h values so that the transformed data were expected to associate with the manipulated levels of skewness and kurtosis. Specifically, when g and h were nonzero,

$$Y = \exp\left(hZ^2/2\right)\cdot[\exp(gZ)-1]/g, \tag{8}$$

where Y is the transformed score and Z is the original normal score. When g was zero,

$$Y = Z\cdot\exp\left(hZ^2/2\right). \tag{9}$$

According to Algina et al. (2005), three types of nonnormal distributions are common in behavioral science. The first type is called a *peaked* (or *kurtosis-based*) distribution, which is characterized by a short (or long) tail of the distribution. Following Algina et al., this study simulated two peaked distributions: (1) $\Upsilon_1 = 0$ and $\Upsilon_2 = 6$ (i.e., $g = 0$ and $h = 0.142$) and (2) $\Upsilon_1 = 0$ and $\Upsilon_2 = 154.84$ (i.e., $g = 0$ and $h = 0.225$). The second type of distribution examined is known as a *skewed* distribution. It is characterized by unequal-length tails between the positive and negative sides of a distribution. In keeping with Algina et al., two skewed distributions were evaluated: (1) $\Upsilon_1 = 2$ and $\Upsilon_2 = 6$ (i.e., $g = 0.76$ and $h = -0.098$; an exponential distribution) and (2) $\Upsilon_1 = 4.90$ and $\Upsilon_2 = 4{,}673.80$ (i.e., $g = 0.225$ and $h = 0.225$). Note that positively (or negatively) skewed distributions often have $\Upsilon_1 > 0$ (or $\Upsilon_1 < 0$), and shorted-tailed (or long-tailed; e.g., t) distributions often have $\Upsilon_2 < 0$ (or $\Upsilon_2 > 0$). The third type of distribution, a *mixed-normal* distribution, appears to be normal to observers, but indeed only 90 % of the observations come from a normal distribution with an *SD* equal to 1.0, and 10 % come from a normal distribution with an *SD* equal to 10. This distribution

has $\Upsilon_1 = 0$ and $\Upsilon_2 = 24.95$, which was found to adversely affect d in Algina et al.'s study.

Factor 2: Total same size (N; three levels). Three levels of N—50, 100, and 300—were simulated, representing small to large sample sizes typically found in behavioral science.

Factor 3: Base rate (b; three levels). *Base rate* is defined as the ratio of sample sizes in Group 1. Following Ruscio and Mullen (2012), the proportions of observations in Group 1 were set at .25, .50, and .75. Hence, the samples sizes could be equal across groups, or one sample could be three times larger than another sample.

Factor 4: *SD* ratio (*SR*; three levels). The SR is the ratio of the *SD*s between two groups, where $SD = \sqrt{\text{Variance}}$. As we noted above, Ruscio and Mullen (2012) stated that *SR*s of $\sqrt{.25}$, $\sqrt{1}$, and $\sqrt{4}$ are common in simulation studies for behavioral and social sciences research. Wilcox (1987) found that *SR*s that exceed 4 are not uncommon in practice. In addition, Grissom and Kim (2001) found that the *SR*s could range from 1.80 to 16.88 in clinical and counseling psychology. Hence, *SR* was set at either 1, 4, and 0.25. The value of 1 assumes that there is homogeneity of variances, and the values of 4 and 0.25 represent violations of the homogeneity assumption that are common in practice.

Factor 5: Population d (δ; five levels). The population values of d were fixed at 0, 0.20, 0.50, 0.80, and 1.50. The levels of 0.20, 0.50, and 0.80 are regarded as small, moderate, and large ESs, respectively (Cohen, 1988). A zero effect (0) and a very large ES (1.5) have been included to evaluate the accuracy of the six ES measures in more extreme conditions. The corresponding population values for r_{pb} are 0, .10, .24, .37, and .60 (Eq. 5), and those for CL and A_w are .50, .56, .64, .71, and .86 (Ruscio, 2008).[2]

The factors were combined to produce a design with 6 × 3 × 3 × 3 × 5 = 810 conditions. Each condition was replicated 10,000 times.

Data generation

For each of the simulation conditions, first, 10,000 random samples of sizes n_1 and n_2 were generated for Y on the basis of a normal distribution, thereby producing the observations in Groups 1 and 2, respectively. Without loss of generality, the population mean (δ_1) and *SD* of the Y scores in Group 1 were set at 0 and 1, respectively. In Group 2, the population *SD* (v) was set at 0.25, 1.00, and 4.00, respectively, and the population mean was fixed at $\delta_2 \cdot \sqrt{[(n_1-1)+(n_2-1)v^2]/(n_1+n_2-2)}$, where δ_2 = (0, 0.20, 0.50, 0.80, 1.50). This process was designed to control the population d at the specified levels. Second, for the first four nonnormal distributions, the generated normal scores were multiplied by the g and h values in Eqs. 8–9, so that they formed a distribution adhering to the manipulated levels of skewness and kurtosis. For the mixed-normal distribution, observations were generated from a uniform distribution $U(0, 1)$. If the observation was less than or equal to .9, then $Y = Z$, else $Y = 10 \cdot Z$. Given the generated observations, the six ES measures were estimated in order to compare their performance. The simulation code was written in Mathematica 10 (Wolfram Research, Inc., 2014), and the code can be found at the homepage https://osf.io/msy3h/.

Evaluation criteria

To evaluate the accuracy of each of the six ESs, percentage bias was used: $bias = \left[\left(\overline{ES}-\delta_t\right)/\delta_t\right]\cdot 100\%$, where $\overline{ES}$ is the mean of the 10,000 ESs obtained in 10,000 simulated samples, and δ_t is the population value of an ES.[3] According to Li, Chan, and Cui (2011), a parameter estimate is considered reasonable when the bias is within ±10 %. Note that the denominator must not be 0 in calculating the bias. Thus, the equation became $bias = \left(\overline{ES}-\delta\right)\cdot 100\%$ when the population ES (δ) was 0.

Results

First of all, the findings of the d_r^* estimates are not included in the following sections because, first, their percentage biases were found to be identical to the biases obtained by d_r. The reason is that the d_r^* estimate was 1.558 (i.e., 1/.642; Eq. 3) times larger than the d_r estimate; hence, when the population value $\delta_t = 1.558 \cdot \delta$ was used for computing the biases of d_r^*, and when the population value, δ, was used for calculating the

[2] According to Ruscio (2008), when the samples sizes are equal for two groups, $r_{pb} = d/\sqrt{d^2+4}, CL = \Phi\left(d/\sqrt{2}\right)$, where Φ is the normal cumulative distribution function, and A_w is the estimator for CL when the data violate the parametric assumption (i.e., normality).

[3] Comparing the robust d_r with the population d (δ) may cause a concern, because d_r is indeed an estimator for the population robust d (i.e., δ_R in Algina et al., 2005) instead of δ. On the other hand, in the present study the original normal scores were generated with a manipulated level of δ, and these scores were transformed to nonnormal data through Eqs. 8 and 9 for the conditions in which the distribution was nonnormal. Hence, the expected standardized mean difference for these scores (either normal or nonnormal) was still δ. In the present study, the accuracy of robust d_r was measured as the bias compared to this value in order to test whether robust d_r could accurately reflect the population d (δ), even when the scores were nonnormally transformed. In Algina et al.'s study, however, the scores were generated on the basis of a manipulated level of δ_R, and hence, the authors directly compared their sample estimates of d_r to the population robust δ_R.

biases of d_r, the two biases became identical. Second, given that the d_r estimate resembled the standardized mean difference metric (*d*-metric) in the conventional *d*, which should be more relevant to researchers in practice, only the results of d_r are reported and discussed in the following sections.

Among the remaining five ES measures, A_w was found to be the most accurate, as is shown in Fig. 1. Of the 810 conditions, 770 (or 95.1 %) yielded a bias within the nominal range of ±10 %. The biases ranged from –13.7 % to 16.2 %, with a mean of –1.5 %, demonstrating excellent accuracy of the A_w measure. Another robust measure, d_r, was found to be appropriate, but it was slightly less accurate than A_w. Of the 810 conditions, 686 (or 84.7 %) produced a bias within ±10 %. The mean of the 810 biases was –3.0 %, which ranged from –54.3 % to 53.1 %. The third ES, *CL*, was generally reasonable. Of the 810 conditions, 648 (or 80 %) yielded a bias within ±10 %. The biases ranged from –31.0 % to 31.1 %, with a mean of –0.6 %.

However, the two parametric-based ESs, *d* and r_{pb}, were not robust to the data violations. The biases ranged from –221.9 % to 184.7 % with a mean of –21.7 % for *d*, and they ranged from –204.2 % to 138.2 % with a mean of –26.4 % for r_{pb}, demonstrating downward-biased estimates of the true ES. Of the 810 conditions, only 258 (or 31.9 %) and 210 (or 25.9 %) produced a bias within ±10 % for *d* and r_{pb}, respectively. The following sections discuss the specific effects of each of the manipulated factors on the ES measures, which are based on the findings shown in Fig. 2.

Effects of the simulated factors on the ES measures

Normal data The four manipulated factors—total samples size (*N*), base rate (*b*), *SD* ratio (*SR*), and population ES (δ)—did not show obvious impacts on *d*, d_r, *CL*, and A_w, with mean biases of 1.1 %, 1.9 %, 5.9 %, and 0.3 %, respectively. This demonstrates that these ESs are appropriate when the data are normal. On the other hand, r_{pb} was slightly less desirable than others, as is evidenced by its largest mean bias (6.9 %). This is because an unbalanced base rate (.25 and .75) would decrease its accuracy. Of the 90 conditions with $b = .25$ and .75, only 30 (or 33.3 %) produced a bias within the nominal range of ±10 %. When $b = .50$, all of the 45 conditions resulted in an appropriate bias. This is understandable, because the parameter $\sqrt{pq}$ in Eq. 4 is influenced by the base rate, which, in turn, affects r_{pb}. In sum, A_w was the most desirable because of its smallest mean bias (0.3 %).

Nonnormal data (peaked) Comparing the five ES measures (i.e., *d*, d_r, r_{pb}, *CL*, and A_w), *CL* was found to be the most accurate and robust to peaked distributions (i.e., $\Upsilon_1 = 0$ and $\Upsilon_2 = 6$; $\Upsilon_1 = 0$ and $\Upsilon_2 = 154.84$). The mean bias was 0.2 %, and the range was [–3.2 %, 3.4 %]. All of the 270 conditions produced a bias within the nominal range of ±10 %, and the MAPE was 1.4 %, showing excellent performance. When δ increased, the bias increased only slightly, but the impact was very minimal. Other factors did not show obvious effects on *CL*.

The second most accurate ES was A_w. Of the 270 conditions, 260 (or 96.3 %) produced a bias within ±10 %. These biases ranged from –11.5 % to 4.0 %, with mean –1.9 %. The MAPE was 2.7 %, which showed good performance. Most of the undesirable results were found under conditions with $b = .25$, $\delta = 1.50$, and $SR = 0.25$, and when $b = .75$, $\delta = 1.50$, $SR = 4$, and $\Theta = 2$. These were conditions of severe violations of the homogeneity of variances and balanced base rate in the present simulation, but the biases were only marginally unacceptable (–10.0 % to –11.5 %). Thus, A_w is regarded as a good estimator for the true ES when the data follow a peaked distribution.

The performance of d_r was found to be comparable to that of A_w. Of the 270 conditions, 264 (or 97.8 %) yielded a bias within ±10 %, and the biases ranged from –11.0 % to 2.5 %, with mean –3.3 %. The MAPE was 3.5 %, demonstrating a good estimate. Most of the unacceptable conditions were observed when $b \neq .50$, $N = 50$, $SR = 1$, and $\Theta = 2$, but these were just marginally beyond the criterion (i.e., –10.2 % to –11.0 %). Hence, d_r is also considered a good ES measure when the data follow a peaked distribution.

Neither *d* nor r_{pb} was an appropriate estimator for the true ES. For *d*, the biases ranged from –35.9 % to 0.2 %, with mean –20.2 %. Of the 270 conditions, only 54 (or 20.0 %) were acceptable. The MAPE was 20.2 %, which was inappropriate. For r_{pb}, the biases ranged from –44.6 % to 1.0 %, with mean –24.6 %. Of the 270 conditions, only 54 (or 20.0 %) were acceptable. The MAPE was 24.6 %, which was undesirable. Thus, the parametric *d* and r_{pb} did not show robustness to violation of normality and should not be reported in practice when the data violate the normality assumption. Because these measures were also inaccurate in the remaining nonnormal data conditions, they will not be discussed in the following sections.

Nonnormal data (skewed) In comparison with other ESs, A_w was the most accurate and robust to skewed distributions (i.e., $\Upsilon_1 = 2$ and $\Upsilon_2 = 6$; $\Upsilon_1 = 4.90$ and $\Upsilon_2 = 4{,}673.80$). The biases ranged from –12.6 % to 16.2 %, with a mean of –0.7 %. Of the 270 conditions, 252 (or 93.3 %) produced a bias within ±10 %. The MAPE was found to be 3.7 %, which is good. The unacceptable biases were found when $b = .25$, $SR \neq 0$, and $\delta > 0.80$. For instance, when $b = .25$, $\delta = 1.5$, and $\Theta = 1$, the biases were slightly larger than 10 % (i.e., 14.0 % to 14.2 %) when $SR = 4$, and they were slightly smaller than –10 % (i.e., –12.0 % to –12.6 %) when $SR = 0.25$. This finding is not surprising, because a larger (or smaller) variance in the more favorable Group 2, with one fourth (i.e., $b = .25$) of the total sample size, may overestimate (or underestimate) the true effect (or mean) in this group, thereby producing an ES that is slightly larger (or smaller) than its true value. Other manipulated factors did not show obvious effects on A_w.

The rescaled robust d_r was generally appropriate, but it was less accurate than A_w. The biases range from –54.3 % to 53.1 %, with a mean of –2.7 %. Of the 270 conditions, 199 (or 73.7 %) resulted in a bias within ±10 %. The MAPE was reasonable (8.7 %). The unacceptable biases were mainly found in conditions when $\delta = 0.20$, $SR \neq 0$, and $\Theta = 1$. For instance, the biases ranged from 23.2 % to 53.1 %, with a mean of 38.3 %, when $\delta = 0.20$, $SR = 4$, and $\Theta = 1$, whereas they ranged from –26.1 % to –54.3 %, with a mean of –38.4 %, when $\delta = 0.20$, $SR = 0.25$, and $\Theta = 1$. This finding is explainable because d_r is the d for the trimmed mean difference over the Winsorized variance, and hence, the observed mean difference does not reflect the true difference, especially when the true difference is less substantial (e.g., $\delta = 0.20$) and the variance ratio is large between the two groups. Or, stated differently, the trimmed d_r contains larger errors due to the (small) true difference being less likely to be identified when a proportion of the observations are discarded, as is done when calculating d_r. When the true ES was larger ($\delta \geq 0.80$), d_r became more stable and accurate.

In the skewed data condition, *CL* did not result in estimates as good as it did in the peaked distribution condition. The biases ranged from –13.0 % to 31.1 %, with a mean of –0.1 %. Of the 270 conditions, 196 (or 72.6 %) resulted in a bias within ±10 %. The MAPE was 8.1 %, which is reasonable. This lower performance was likely due to the fact that the unbalanced tails decreased the accuracy of the normality-based cumulative distribution function (Φ) when *CL* was computed in Eq. 6.

Nonnormal data (mixed-normal) Similar to the results obtained from the skewed distributions, both A_w and d_r were appropriate when data were mixed-normal, but A_w slightly outperformed d_r. Regarding A_w, the biases ranged from –13.7 % to 0.9 %, with a mean of –3.3 %. Of the 135 conditions, 123 (or 91.1 %) were within the nominal range of ±10 %. The MAPE was 3.5 %, which is appropriate. The 12 unacceptable conditions were found when $b = .25$, $SR = 0.25$, and $\delta \geq 0.80$, and when $b = .75$, $SR = 4$, and $\delta \geq 0.80$, but the biases were just slightly beyond the criterion of ±10 % (i.e., –10.4 % to –13.7 %). Hence, A_w is regarded as robust to the mixed-normal distribution, even when the variance ratio and base rate are unbalanced between the two groups.

For d_r, the biases ranged from –15.3 % to 0.3 %, with a mean of –7.7 %. Of the 135 conditions, 89 (or 65.9 %) yielded a bias within ±10 %. The MAPE was 7.7 %, which is reasonable. When $\delta = 0$, all of the biases in the 27 conditions were highly desirable. When $\delta \geq 0.20$, the undesirable biases occasionally appeared when $b = .25$ or $b = .75$ and other factors were held constant. Specifically, of the 72 conditions, 38 (or 52.8 %) resulted in a bias between –10 % and –15.3 %, which showed a slight downward bias greater than the acceptable criterion of ±10 %. In sum, A_w is more accurate than d_r, although both ESs are deemed reasonable.

Fig. 2 Percentage biases of the effect sizes listed by the manipulated factors. *ES* is an effect size that includes d (Cohen's d), d_r (rescaled robust d), r_{pb} (point-biserial correlation), *CL* (common-language ES), and A_w (nonparametric estimator for *CL*). *SR* is the *SD* ratio, n is the total sample size, θ is the data distribution, δ is the true ES value in the d-metric, and b is the base rate

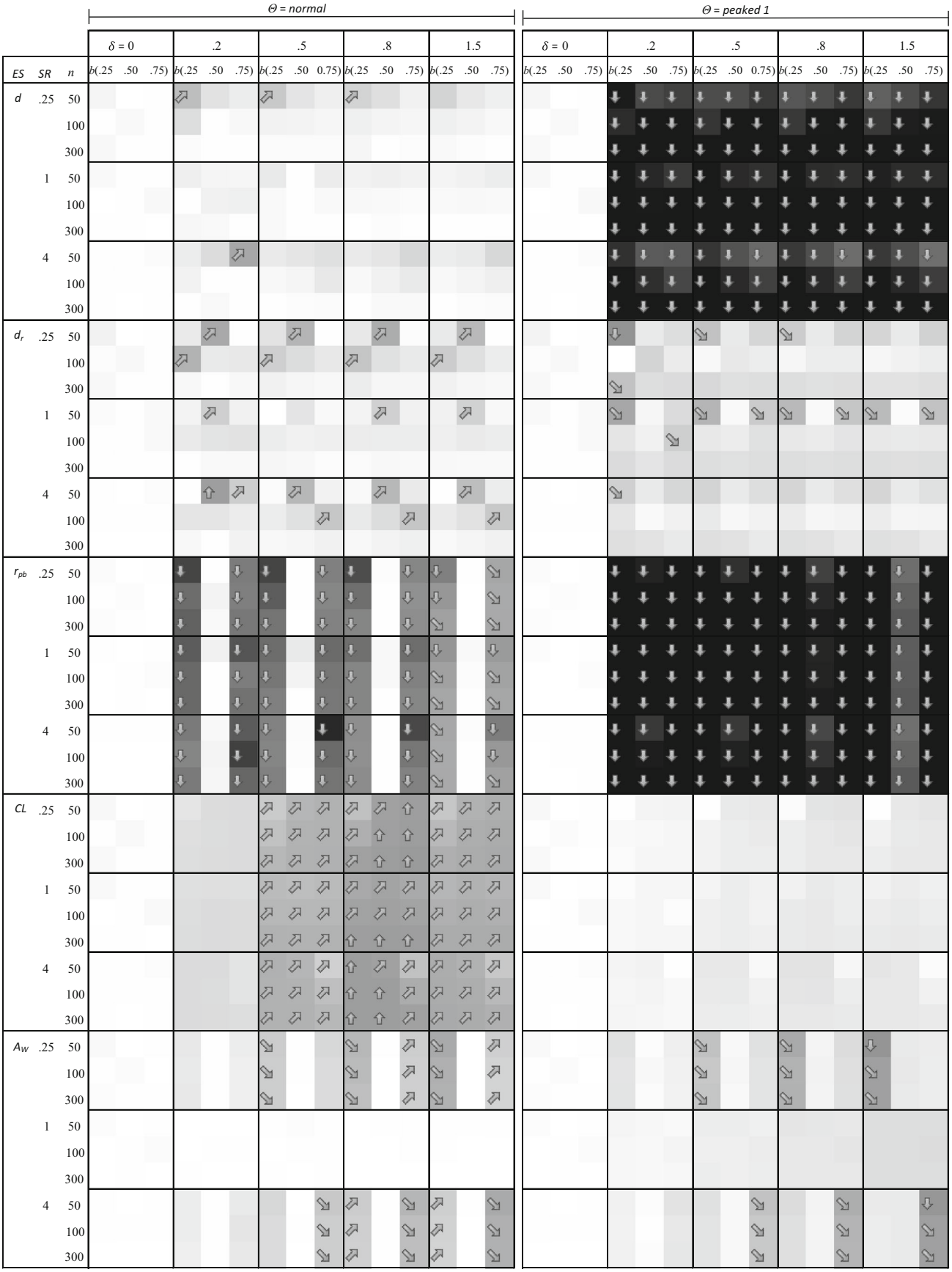
CL did not provide desirable ES estimates. The biases ranged from –20.8 % to 0.4 %, with mean –9.9 %. Of the 135 conditions, only 60 (or 44.4 %) resulted in an acceptable bias. The MAPE was 9.9 %, which was just smaller than the criterion of 10 %. This showed that *CL* was highly sensitive to a mixed-normal distribution even when only 10 % of the observations followed a normal distribution with a larger variance than the remaining 90 %. Thus, *CL* is not recommended in general.

Conclusion and discussion

This article evaluated the performance of six ES measures when data violated the assumptions of normality and homogeneity of variances, circumstances that are common in behavioral science research. The results showed that both A_w and d_r were generally robust to the violations of these assumptions. Specifically, A_w slightly outperformed d_r especially when the data followed a skewed distribution (i.e., exponential distribution; $\Upsilon_1 = 2$ and $\Upsilon_2 = 6$) and a mixed-normal distribution (i.e., $\Upsilon_1 = 0$ and $\Upsilon_2 = 24.95$). The conventional d and r_{pb}, however, were not robust to these violations, and hence, they should not be reported, or should at least be interpreted with caveats in practice. The following sections discuss the practical implications of using A_w and d_r, and also provide guidelines for researchers and practitioners to report and interpret ESs when reporting their findings.

Interpreting A_w

Given that A_w was found to be an accurate estimator for *CL* in this study, researchers are encouraged to report and interpret A_w directly, especially when their data violate the assumptions of normality and the homogeneity of variances. The conventional rule of thumb for small, moderate, and large ESs in d (i.e., 0.20, 0.50, and 0.80) can be easily converted to the A_w-metric (i.e., .56, .64, .71; see the equations in note 2). According to Ruscio (2008), A_w communicates ES in a common-language way, so that laypersons can understand the meaning of superiority in one group over another. For example, if a cognitive psychology researcher finds that the observed


Θ = normal
Θ = peaked 1
δ = 0
.2
.5
.8
1.5
ES
SR
n
b(.25 .50 .75)
b(.25 .50 0.75)
d
d_r
r_{pb}
CL
A_W
.25
1
4
50
100
300

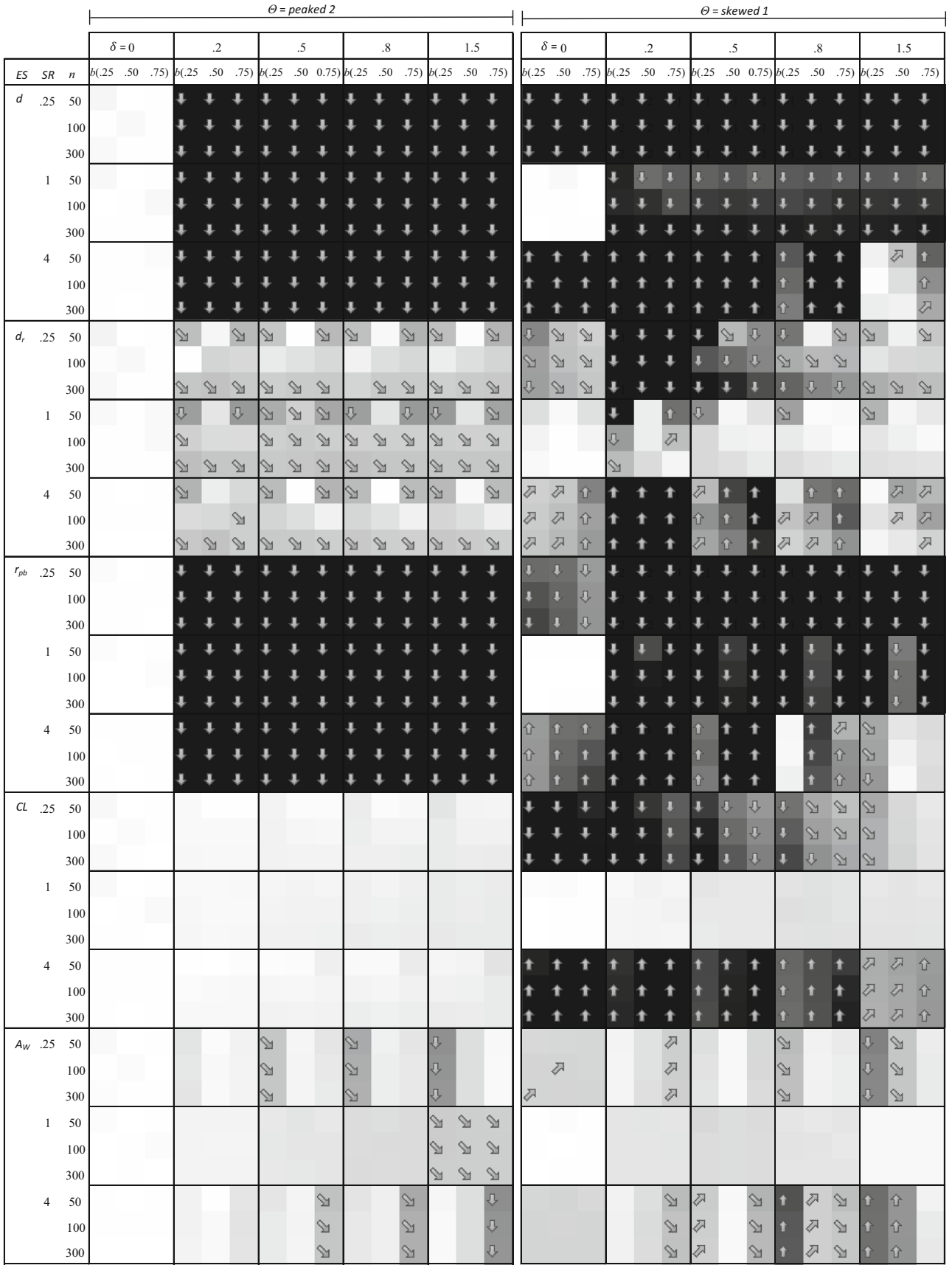


Fig. 2 (continued)

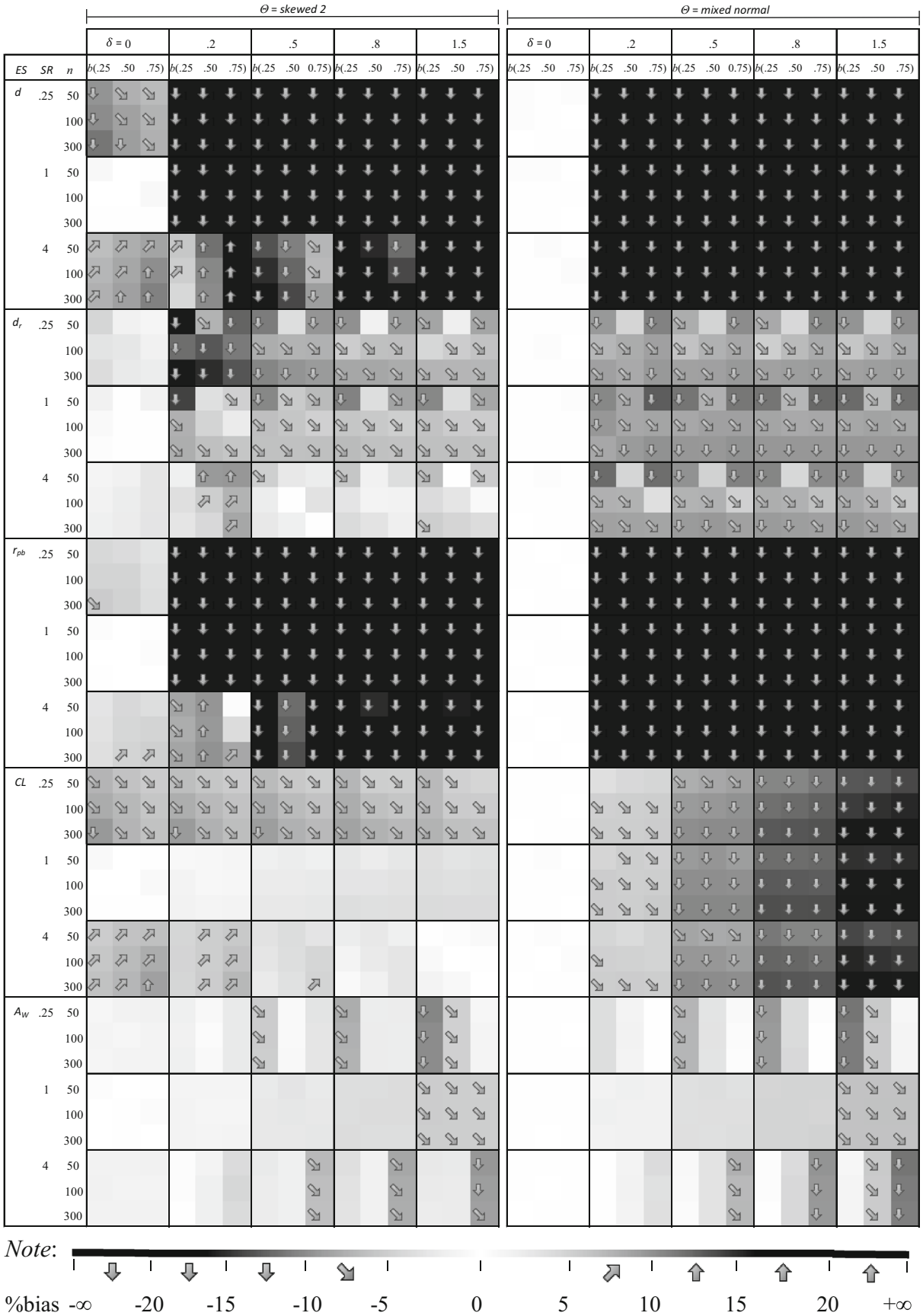


Fig. 2 (continued)

A_w is .71 when examining the difference in typing speed between female and male participants, then the researcher can conclude that there is a 71 % chance that a randomly selected female participant would possess a faster typing speed than a randomly selected male participant. This is regarded as a large ES between the two groups.

Interpreting the d-metric ES through A_w or d_r

In light of the popularity of d, researchers and practitioners are more familiar with the interpretation of d in real-world research. The prevalence notwithstanding, d was found to be inaccurate when the normality and homogeneity-of-variances assumptions were violated in this study, thereby severely affecting the accuracy of d in evaluating the true ES in the research literature. This article provides two alternative estimators for the true ES—A_w and d_r—which are more robust than the conventional d in Eq. 1.

A hypothetical data set was simulated to demonstrate the interpretative procedure with manipulated factors: Θ = exponential distribution, base rate = .50, population ES = 0.50, SR = 0.25, and total sample size = 50, producing Group 1 = {–0.831, –0.745, –0.735, –0.716, –0.510, –0.509, –0.471, –0.448, –0.378, –0.283, –0.041, 0.024, 0.151, 0.174, 0.193, 0.299, 0.346, 0.523, 0.808, 0.854, 1.050, 1.608, 2.491, 4.536, 4.970}, and Group 2 = {0.156, 0.176, 0.198, 0.200, 0.210, 0.220, 0.227, 0.228, 0.237, 0.269, 0.277, 0.307, 0.311, 0.351, 0.355, 0.361, 0.399, 0.437, 0.464, 0.483, 0.503, 0.519, 0.565, 0.616, 0.674}. When one evaluates the ES between Groups 2 and 1 (i.e., Group 2 minus Group 1), the observed d becomes –0.135 [i.e., $(.350-.494)/\sqrt{[(25-1)\cdot(.022)+(25-1)\cdot(2.279)]/(25+25-2)}$ Eq. 1], meaning that the ES is small, and the observations are smaller in Group 2 than Group 1. However, this interpretation is highly inaccurate because the true ES is indeed 0.50. Reporting d = –0.135 causes a serious problem in that it leads to an inaccurate interpretation of the actual observed ES between the two groups.

To improve the accuracy, one can use d_r. In this example, d_r is 0.392 [i.e., $.642\cdot(.328-.083)/\sqrt{[(25-1)\cdot(.012)+(25-1)\cdot(.310)]/(25+25-2)}$ Eqs. 2 and 3], which is closer to the true ES of 0.50. The interpretation for 0.392 will be a small-to-moderate ES with the observations larger in Group 2.

The most accurate estimator of the true ES in this example is A_w, which is equal to .6416 [i.e., 401/(25·25) Eq. 7]. Converting A_w to the d-metric, the value becomes 0.51 (i.e., $d_A = \sqrt{2}\cdot\Phi^{-1}(.6416)$; equation in note 1), which is almost identical to the true ES of 0.50. Thus, researchers and practitioners are encouraged to compute A_w and convert it to the d-metric for interpreting the ES, especially when the data violate the normality and homogeneity-of-variances assumptions (with the restrictions that the data are not mixed-normal and the observed A_w is neither 0 nor 1).[4]

Research scenarios for A_w or d_r

In addition to the empirical evidence supporting the appropriateness of A_w and d_r researchers should also consider which type of ES makes the most sense to report in their particular research domain. In particular, A_w and d_r express very different kinds of effects, and researchers should choose between them on the basis of their meaningfulness within the research domain. Take, for example, a researcher interested in comparing the difference in communication skills between female and male college students. If the researcher is interested in presenting a magnitude that reflects the difference between the two groups, and finds that the data do not follow the conventional parametric assumptions (i.e., normality and homogeneity of variances), the researcher should report d_r (e.g., 0.50). This choice still accurately presents that the female students, on average, score 0.50 *SD*s higher than the male students in communication skills. On the other hand, if the researcher intends to present how likely a randomly selected female student would be to outperform a randomly selected male student (or vice versa) from the same data set, the researcher should report A_w (e.g., .64). This choice accurately presents that there is a 64 % chance that a randomly chosen female student would possess better communication skills than a randomly chosen male student.

Application of A_w and d_r in meta-analysis

It is also important to note a potential application of A_w and d_r in meta-analysis. A common research interest in behavioral research involves the summary or meta-analysis of a subgroup difference (e.g., males vs. females) in a numeric variable (e.g., cognitive ability). Depending on the research interest of a meta-analyst, the meta-analyst can either provide a summary of the standardized mean differences (d-metric) or the probability of superiorities (A_w-metric) in a research domain. If one is interested in pooling the d-metric statistics, the ds are

[4] Note that the transformation from A_w to d_r is not linear [i.e., $d_A = \Phi^{-1}(A_W)/\sqrt{(p_1 s_1^2 + p_2 s_2^2)/s_1^2 + s_2^2}$, where p_i is the proportion of observations and s_i^2 is the variance for group i = 1, 2; Ruscio, 2008]. Hence, the bias of the converted d_A could be different from that of the original A_w and could be better or poorer than that of the d_r estimate. Regarding the performance of d_A as compared to d_r, the biases of d_A ranged from –115.78 % to 101.13 %, with a mean of –1.80 %, when distributions were normal, peaked 1, peaked 2, skewed 1, or skewed 2. Comparatively, the percentage biases of the d_r estimates ranged from –120.50 % to 107.93 %, with a mean of –1.97 %, which is slightly less accurate than the d_A estimates. However, this advantage diminished when the distribution was mixed-normal. The biases ranged from –21.69 % to 0.29 %, with a mean of –12.71 %, for d_A, and they ranged from –15.26 % to 0.32 %, with a mean of –7.58 %, for d_r. Moreover, an observed value for A_w should be neither 0 nor 1, because the transformed d_A would become negative or positive infinity.

usually either directly found in published studies or can be estimated from the descriptive statistics provided in these studies. For studies in which the parametric assumptions are violated, if d_r is reported in these studies, it can be directly plugged into the mean d, because d_r is a robust estimator for the true population value.

A potential issue with this approach is that d_r may not have been widely employed in the existing literature since its development in Algina et al. (2005). If that is the case, one can first search for the A_w statistic, which may either be reported in published studies or calculated from the Mann–Whitney U statistic, which is a popular nonparametric statistical-significance test that is an alternative to the conventional independent-samples t test—that is, $A_w = (n_1n_2 - U)/n_1n_2$, where $U = [\#(\boldsymbol{p} > \boldsymbol{q}) + .5\#(\boldsymbol{p} = \boldsymbol{q})]$ is the Mann–Whitney U statistic. Next, one can transform an observed A_w to d_A—that is, $d_A = \Phi^{-1}(A_W)/\sqrt{\left(p_1s_1^2 + p_2s_2^2\right)/\left(s_1^2 + s_2^2\right)}$, where p_i is the proportion of observations and s_i^2 is the variance for group i = 1, 2, and Φ^{-1} is the inverse normal cumulative distribution function (Ruscio, 2008). Hence, d_A is a robust estimator for the true population value, which can be used for pooling the ds in meta-analysis. On the other hand, if one attempts to pool the A_w-metric statistics, one can transform the published ds (and d_rs) into the A_w statistics—that is, $A_{W_d} = \Phi\left(d \cdot \sqrt{\left(p_1s_1^2 + p_2s_2^2\right)/\left(s_1^2 + s_2^2\right)}\right)$, with an assumption that the data of the reported ds met the parametric assumptions in the original studies.

Future directions

A first direction for future research involves extending the framework of robust ES measures (e.g., A_w, d_r) in the two-independent-samples case to the more general univariate and multivariate analysis-of-variance (ANOVA) scenarios that involve single or multiple independent variables with more than two groups and multiple DVs. Ruscio and Gera (2013) have generalized A_w in the univariate ANOVA scenario, but their study did not systematically evaluate the benefits of A_w in comparison with the conventional ESs—eta-squared, partial eta-squared, and omega-squared—used in ANOVA. Moreover, no study has discussed the generalization of d_r in the ANOVA framework.

Second, future research can examine the performance of the CIs surrounding each of the ESs (see the Appendix). For example, Ruscio and Mullen (2012) and Algina et al. (2005) have examined the performance of bootstrap CIs for A_w and d_r, respectively, in a two-independent-samples case. However, these studies did not investigate the bootstrap CIs for these robust ESs in more general univariate and multivariate ANOVA scenarios. Thus, additional studies will be needed to further examine the robustness of these ESs as well as their CIs in more general cases.

Appendix: Confidence intervals for the ESs

Two different types of CIs (parametric and nonparametric) can be used to estimate the sampling error for the ESs in this study. The parametric CIs often require a mathematical proof or analytic equation that is usually based on the parametric assumptions (e.g., normality), and the CIs are typically symmetric in terms of the upper and lower limits surrounding a point estimate. On the other hand, the nonparametric CIs do not depend upon these assumptions, and the bootstrap (resampling) procedure has been frequently used to estimates these CIs. Because researchers and practitioners usually report a 95 % CI in practice, the following equations are reported on the basis of this percentage.

Parametric CIs

CI for *d* According to Algina et al. (2005), one can make use of the noncentral t distribution in order to construct a 95 % CI for either d or d_r. Regarding d, first, find the lower and upper limits from the noncentral t distribution, conditional on two parameters: (a) degrees of freedom (i.e., $df = n_1 + n_2 - 2$) and (b) a noncentrality parameter [i.e., $\lambda = d\sqrt{n_1n_2/(n_1 + n_2)}$]. Second, when these two parameters (i.e., df and λ) are found, identify the lower limit (i.e., 2.5 % percentile; l_λ) and upper limit (97.5 % percentile; u_λ) for λ from the noncentral t distribution. Third, transform these lower and upper limits to the d-metric—that is, $l_d = l_\lambda \sqrt{n_1 + n_2/n_1n_2}$ and $u_d = u_\lambda\sqrt{(n_1 + n_2)/n_1n_2}$.

CI for d_r According to Algina et al. (2005), the above-mentioned procedure can be applied to d_r replaced by robust parameters. First, the two parameters for obtaining the lower and upper limits (i.e., l_{d_r} and u_{d_r}) for the robust noncentrality parameter (λ_r) become (a) degrees of freedom, $df_r = h_1 + h_2 - 2$, where h_i is the number of observations remaining after trimming for group i = 1, 2, and (b) noncentrality parameter $\lambda_r = d_r\sqrt{h_1h_2/(h_1 + h_2)}$. After obtaining the l_{d_r} and u_{d_r} from the noncentral t $d$istribution, transform them into the d-metric—that is, $l_{d_r} = l_{\lambda_r} \sqrt{(h_1 + h_2)/h_1h_2}$ and $u_{d_r} = u_{\lambda_r}\sqrt{(h_1 + h_2)/h_1h_2}$.

CI for r_{pb} According to Tate (1955), a CI for r_{pb} can be estimated by $CI_{r_{pb}} = r_{pb} \pm 1.96\sqrt{\frac{r_{pb}^2+2p_1(1-p_1)\left(2-3r_{pb}^2\right)}{4np_1(1-p_1)}\left(1-r_{pb}^2\right)^2}$, where p_1 is the proportion of scores in Group 1, and n is the total sample size.

CIs for CL and A_w According to Ruscio and Mullen (2012), there are at least seven different analytic procedures for the CI surrounding *CL* or A_w. Among them, Cliff's (1993) CI procedure is regarded as a popular CI construction method—that is,

$$CI_{CL} = CL \pm 1.96\sqrt{\frac{n_1^2\sum_{i=1}^{n_1}\left(D_{i\cdot}-D\right)^2 + n_2^2\sum_{j=1}^{n_2}\left(D_{\cdot j}-D\right)^2 - \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(D_{ij}-D\right)^2}{n_1n_2(n_1-1)(n_2-1)}},$$

where $D_{ij} = sign(Y_i - Y_j)$ refers to the dominance score, in which $D_{ij} = 1$ if $Y_i > Y_j$, $D_{ij} = -1$ if $Y_i < Y_j$, and $D_{ij} = 0$ if $Y_i = Y_j$, $D = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}D_{ij}/n_1n_2$, and $D_{i\cdot}$ is the marginal mean of the dominance scores in Group 2, and $D_{\cdot j}$ is the marginal mean of the dominance scores in Group 1.

Given that A_w is the robust estimator for *CL*, the same equation can be used—that is, $CI_{A_W} = A_W \pm 1.96 \times \sqrt{\frac{n_1^2\sum_{i=1}^{n_1}\left(D_{i\cdot}-D\right)^2 + n_2^2\sum_{j=1}^{n_2}\left(D_{\cdot j}-D\right)^2 - \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(D_{ij}-D\right)^2}{n_1n_2(n_1-1)(n_2-1)}}$.

Nonparametric CIs

Another popular CI construction procedure involves using a nonparametric (or nonanalytic) procedure. In the existing literature, the bootstrap CI has been found to be appropriate for the construction of the CIs, especially when the data violate the parametric assumptions, the analytic equation for the CI is difficult to derive, and the confidence widths are asymmetric for the upper and lower ends (Beasley & Rodgers, 2009). Generally, the bootstrap procedure first resamples a data set with replacement to form B (e.g., 5,000) numbers of bootstrap samples, and each of them has the same sample size as the original dataset. Second, an ES estimate (i.e., d, d_r^*, d_r, r_{pb}, *CL*, or A_w) is computed for each of the bootstrap samples, thereby producing B numbers of the bootstrap ESs (δ_Bs). Third, these bootstrap ESs are rank-ordered in ascending order to derive a sampling distribution. Consequently, the rank-ordered bootstrap ESs are sufficient for constructing three different types of bootstrap CIs—the bootstrap standard interval (BSI), bootstrap percentile interval (BPI), and bootstrap bias-corrected and accelerated interval (BCaI)—that a researcher can choose from. Generally, BSI constructs a 95 % CI by BSI = $\delta_t \pm 1.96s_B$, where s_B is the *SD* of the B = 5,000 bootstrap ESs, and δ_t is any one of the five ESs examined in the present study. Regarding BPI, the 2.5 and 97.5 percentiles of the B = 5,000 bootstrap ESs are extracted—that is, BPI = $[\delta_B(l), \delta_B(u)]$, l = 2.5 percentile rank, and u = 97.5 percentile rank. The BCaI is regarded as a bias-adjusted BPI, in which BCaI = $[\delta_B(l'), \delta_B(u')]$, where $l' = B{\cdot}\Phi\left\{i + \frac{a+z_{1-(\alpha/2)}}{1-b\left[a+z_{1-(\alpha/2)}\right]}\right\}$, and $u' = B{\cdot}\Phi\left\{i + \frac{a-z_{1-(\alpha/2)}}{1-b\left[a-z_{1-(\alpha/2)}\right]}\right\}$. The first correction parameter in the equation, $a = \Phi^{-1}\{\#[\delta_b(b) < \delta_t]/B\}$ is a correction factor that accounts for the overall bias (e.g., skewness) of the bootstrap ESs (δ_b) that deviate from the original ES estimate (δ_t), where Φ^{-1} is the normal inverse cumulative function distribution, and $\#[\delta_b(b) < \delta_t]$ is the count function that counts the number of the bootstrap ESs below the estimate δ_t in the original data set. The second correction parameter (b) corrects for the rate of change of δ_t with respect to its true parameter value—that is, $b = \sum_{k=1}^{K}[\delta_t(.) - \delta_t(k)]^3/6\{\sum_{k=1}^{K}[\delta_t(.) - \delta_t(k)]^2\}^{3/2}$, where $\delta_t(k)$ is the jackknife value of the ES estimate obtained by removing the kth row of the original data set, and $\delta_t(.)$ is the mean of the n jackknife ES estimates.

References

Algina, J., Keselman, H. J., & Penfield, R. D. P. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10,* 317–328. doi:10.1037/1082-989X.10.3.317

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

Beasley, W. H., & Rodgers, J. L. (2009). Resampling methods. In R. E. Millsap & A. Maydeau-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 362–386). Thousand Oaks, CA: Sage.

Brown, R. A., Evans, D. M., Miller, I. W., Burgess, E. S., & Mueller, T. I. (1997). Cognitive-behavioral treatment for depression in alcoholism. *Journal of Consulting and Clinical Psychology, 65,* 715–726. doi:10.1037/0022-006X.65.5.715

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

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

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

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to

moderate sized samples. *Psychological Methods, 7,* 485–503. doi:10.1037/1082-989X.7.4.485

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141,* 2–18. doi:10.1037/a0024338

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79,* 314–316. doi:10.1037/0021-9010.79.2.314

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6,* 135–146. doi:10.1037/1082-989X.6.2.135

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.

Hogarty, K. Y., & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Hsu, L. M. (2004). Biases of success rate differences shown in binomial effect size displays. *Psychological Methods, 9,* 183–197. doi:10.1037/1082-989X.9.2.183

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.

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. doi:10.1348/2044-8317.002007

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods, 11,* 386–401. doi:10.1037/1082-989X.11.4.386

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

Ruscio, J., & Gera, B. L. (2013). Generalizations and extensions of the probability of superiority effect size estimator. *Multivariate Behavioral Research, 48,* 208–219. doi:10.1080/00273171.2012.738184

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research, 47,* 201–223. doi:10.1080/00273171.2012.658329

Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research: A review and a new index. *Methodology, 8,* 1–11. doi:10.1027/1614-2241/a000034

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.

Tate, R. F. (1955). Applications of a correlation models for biserial data. *Journal of the American Statistical Association, 272,* 1078–1095.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistic of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25,* 101–132. doi:10.3102/10769986025002101

Weisz, J. R., Weiss, B., Han, S. S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin, 117,* 450–468. doi:10.1037/0033-2909.132.1.132

Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology, 38,* 29–60. doi:10.1146/annurev.ps.38.020187.000333

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Elsevier Academic Press.

Wolfram Research, Inc. (2014). *Mathematica (Version 10.0)*. Champaign, IL: Author.