Original Article

Research Synthesis Methods

Received 12 October 2015, Revised 07 April 2016, Accepted 13 May 2016 Published online 15 September 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1218

Estimation of the biserial correlation and its sampling variance for use in meta-analysis

Perke Jacobs[a]* and Wolfgang Viechtbauer[b]

Meta-analyses are often used to synthesize the findings of studies examining the correlational relationship between two continuous variables. When only dichotomous measurements are available for one of the two variables, the biserial correlation coefficient can be used to estimate the product–moment correlation between the two underlying continuous variables. Unlike the point-biserial correlation coefficient, biserial correlation coefficients can therefore be integrated with product–moment correlation coefficients in the same meta-analysis. The present article describes the estimation of the biserial correlation coefficient for meta-analytic purposes and reports simulation results comparing different methods for estimating the coefficient's sampling variance. The findings indicate that commonly employed methods yield inconsistent estimates of the sampling variance across a broad range of research situations. In contrast, consistent estimates can be obtained using two methods that appear to be unknown in the meta-analytic literature. A variance-stabilizing transformation for the biserial correlation coefficient is described that allows for the construction of confidence intervals for individual coefficients with close to nominal coverage probabilities in most of the examined conditions. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: biserial correlation; bias; sampling variance; variance-stabilizing transformation; confidence intervals; meta-analysis

1. Introduction

Meta-analyses are frequently used to synthesize the results of studies that have examined the association between two variables of interest (e.g., Cooper, 2009; Hunter and Schmidt, 2004; Rosenthal, 1991). When interest lies in the (linear) relationship between two variables that reflect a continuum (e.g., depression and anxiety), the strength and the direction of the relationship are usually expressed in terms of the Pearson product–moment correlation coefficient (e.g., Borenstein, 2009; Fisher, 1915). A set of raw correlation coefficients extracted from a collection of studies examining the same relationship can then be aggregated and contrasted with standard meta-analytic procedures (e.g., Shadish and Haddock, 2009). Alternatively, the raw correlation coefficients can first be transformed with Fisher's *r*-to-*z* transformation (Fisher, 1921) before applying such procedures.

Ideally, all studies providing evidence about the association between the two variables of interest have measured the two variables in their continuous form and reported the resulting sample Pearson product–moment correlation coefficients, which can then be directly used for the meta-analysis. However, studies sometimes report results with one or both variables of interest artificially dichotomized, despite known and frequently repeated arguments against this practice (e.g., Cohen, 1983; MacCallum *et al*., 2002; Maxwell and Delaney, 1993). Such dichotomization may occur as part of the measurement process, when a variable that reflects a continuum was assessed dichotomously, or during the analysis stage, when a variable that was actually measured on a continuum was dichotomized prior to the analysis. Dichotomization may be carried out "adaptively" (e.g., at a cutoff point that depends on the particular sample observed, such as when dichotomizing a variable at the sample median) or using a "hard" cutoff (e.g., at a predefined and absolute value that may be considered to be of clinical and/or practical relevance).

[a]*Max Planck Institute for Human Development, Berlin, Germany*
[b]*Maastricht University, Maastricht, The Netherlands*
**Correspondence to: Perke Jacobs, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany.*
E-mail: jacobs@mpib-berlin.mpg.de

If both continuous variables have been dichotomized, study authors may report sufficient information such that the counts for the resulting 2×2 table of cell frequencies can be reconstructed. One can then calculate the tetrachoric correlation coefficient (Pearson, 1900, 1913), which provides an estimate of the true product–moment correlation between the two underlying continuous variables, assuming that the continuous variables follow a bivariate normal distribution.

If only one of the two continuous variables has been dichotomized, study authors are likely to report summary statistics (i.e., means and standard deviations) for the continuous variable, stratified by the two levels of the dichotomous variable. Based on this information, one can calculate the biserial correlation coefficient, which again provides an estimate of the true underlying product–moment correlation between the two continuous variables (Pearson, 1909a; Soper, 1914; Tate, 1955a, 1955b).

It is important to note that neither the phi coefficient (Boas, 1909; Pearson, 1900; Pearson, 1909b; Yule, 1912), which can also be calculated based on the 2×2 table data, nor the point-biserial correlation coefficient (Lev, 1949; Tate, 1954), which can also be calculated based on the stratified summary statistics, provides appropriate estimates of the relationship between the two underlying continuous variables. These coefficients do not account for the fact that one or both variables have been dichotomized artificially and therefore do not estimate the correlation of the underlying bivariate normal distribution. Therefore, if information from these different study types is to be combined in a single meta-analysis, only product–moment correlations, tetrachoric correlations, and biserial correlations provide a set of coefficients that are inherently and logically comparable from a statistical point of view.

Methods to calculate the tetrachoric correlation coefficient (and its associated sampling variance) have been described in the literature in much detail. In particular, given the 2×2 table of cell frequencies, the maximum likelihood estimate can be computed using iterative numerical procedures (e.g., Brown, 1977; Hamdan, 1970; Olsson, 1979; Pearson, 1900; Pearson, 1913; Tallis, 1962). Alternatively, if only the odds ratio based on the 2×2 table is reported, one can use one of several approximations to estimate the tetrachoric correlation coefficient (e.g., Becker and Clogg, 1988; Bonett and Price, 2005; Digby, 1983; Pearson, 1900).

In the present paper, our focus is on the biserial correlation coefficient and its associated sampling variance. Despite the fact that the biserial correlation is occasionally described in the meta-analytic literature (Hunter and Schmidt, 1990; Hunter and Schmidt, 2004; Lipsey and Wilson, 2001), its estimation based on stratified summary statistics has not been fully explained in this context. Moreover, methods for estimating the sampling variance of a biserial correlation coefficient presented in the literature rely on various simplifications or approximations that may not be sufficiently accurate for practical use. At the same time, a derivation of the asymptotic sampling variance of the biserial correlation coefficient (Soper, 1914) appears to have remained unnoticed in the meta-analytic literature.

The goal of the present paper is to describe how the biserial correlation coefficient can be accurately estimated based on the information that one would typically find reported when one of the two variables of interest has been artificially dichotomized by the study authors. We then describe various methods for estimating the sampling variance of the biserial correlation coefficient, which leads to a discussion of different methods for constructing confidence intervals based on biserial data. The results of a Monte Carlo simulation study comparing the accuracy of these different procedures will then be presented. Finally, the article concludes with an illustration of a meta-analysis that combines information from studies providing product–moment correlation coefficients and studies providing biserial and tetrachoric correlations and a discussion of some general considerations when carrying out such a meta-analysis.

2. The biserial correlation coefficient

A large collection of methods for estimating population correlation coefficients has been described in the statistical literature. Most famously, the Pearson product–moment correlation coefficient ρ between two continuous variables, say X and Y, is defined as follows:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{1}$$

where σ_{XY} denotes the population covariance between X and Y and σ_X and σ_Y denote the respective population standard deviations. In a sample of size n from the bivariate distribution of X and Y, the product–moment correlation coefficient is estimated analogously by the sample correlation coefficient, given by

$$r = cor(x, y) = \frac{cov(x, y)}{s_x s_y}, \tag{2}$$

where $cov(x, y)$ denotes the sample covariance between the observed x and y values and s_x and s_y denote their respective sample standard deviations.

Table 1. Example data.

No	x	y	x_b	Summary statistics
1	20	3	1	
2	31	3	1	
3	18	4	1	
4	22	5	1	
5	30	6	1	$\overline{y}_1 = 4.7$
6	16	4	1	$s_1 = 1.337$
7	28	7	1	$n_1 = 10$
8	24	6	1	
9	23	5	1	
10	27	4	1	
11	1	3	0	
12	4	5	0	
13	8	1	0	
14	15	5	0	
15	9	2	0	$\overline{y}_0 = 3.6$
16	11	4	0	$s_0 = 1.578$
17	11	6	0	$n_0 = 10$
18	6	4	0	
19	8	2	0	
20	4	4	0	
		$r = 0.44$	$r_b = 0.46$	$r_{pb} = 0.37$

For example, consider a population in which $\sigma_{XY} = 9$, $\sigma_X = 9$, and $\sigma_Y = 2$, implying $\rho = 0.5$. Suppose a sample of size $n = 20$ is taken from this population, yielding the data as shown in the columns labeled x and y in Table 1. For these data, we find $cov(x, y) = 6.347$, $s_x = 9.446$, and $s_y = 1.531$. Based on this information, ρ is estimated to be $r = \frac{6.347}{9.446 \times 1.531} = 0.44$.

Estimation of ρ becomes more complicated when one of the observed variables, say x, has been dichotomized at a certain cutoff (here denoted by c) to yield a dichotomous or binary variable x_b, taking on one of only two possible values, say 0 and 1. Data of this format essentially consist of the observed values of variable y within the two groups defined by the levels of variable x_b. Study authors are then likely to report summary statistics (i.e., means and standard deviations) of variable y stratified by the group variable x_b. Alternatively (or in addition), authors may report the results from an independent sample t-test for the test of the null hypothesis H_0: $\mu_1 = \mu_0$, where μ_1 and μ_0 denote the true means of variable Y within the two levels of variable x_b.

For the application considered here, interest lies in the underlying correlation between the non-dichotomized variable X and variable Y, ρ, which is estimated by the biserial correlation coefficient. The steps to calculate the biserial correlation from the stratified summary statistics are as follows:

First, we calculate the *standardized mean difference* between the two groups (i.e., Cohen's d) using

$$d = \frac{\overline{y}_1 - \overline{y}_0}{\sqrt{\frac{(n_1-1)s_1^2 + (n_0-1)s_0^2}{n_1+n_0-2}}}, \tag{3}$$

where $\overline{y}_1$ and $\overline{y}_0$ denote the sample means of the two groups, s_1 and s_0 denote the sample standard deviations, and n_1 and n_0 denote the group sizes. If only the group sizes and the test statistic for the independent samples t-test, here denoted by t, are reported by the study authors, then the value of d can also be computed with

$$d = t\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}. \tag{4}$$

Consider again the data in Table 1. The column labeled as x_b illustrates one possible way of dichotomizing variable x, using a cutoff value of $c = 15.5$, corresponding to the median of variable x. Therefore, the groups are of sizes $n_1 = n_0 = 10$. For these data, we find $\overline{y}_1 = 4.7$, $\overline{y}_0 = 3.6$, $s_1 = 1.337$, and $s_0 = 1.578$. The standardized mean difference can then be computed with

$$d = \frac{4.7 - 3.6}{\sqrt{\frac{(10-1)1.337^2 + (10-1)1.578^2}{10+10-2}}} = 0.75.$$

The results from an independent samples t-test (i.e., $t(\text{df} = 18) = 1.68$, $p = 0.11$) can be used to compute the same value with

$$d = 1.68\sqrt{\frac{1}{10} + \frac{1}{10}} = 0.75.$$

In the next step, the standardized mean difference is transformed into the *point-biserial correlation coefficient*, which is simply the value one would obtain when applying the equation for the product–moment correlation coefficient to variables x_b and y. Based on d, this value can be computed with

$$r_{pb} = \frac{d}{\sqrt{d^2 + h}}, \tag{5}$$

where h is defined as $h = \frac{m}{n_1} + \frac{m}{n_0}$ and $m = n_1 + n_0 - 2$.

Alternatively, if the sample means of the two groups are given, but only s_y, the standard deviation of the y scores (i.e., not separately for the two groups), we can also directly calculate r_{pb} with

$$r_{pb} = \left(\frac{\overline{y}_1 - \overline{y}_0}{s_y}\right)\sqrt{\frac{npq}{n-1}}, \tag{6}$$

where p and q are defined as $p = \frac{n_1}{n}$ and $q = 1 - p = \frac{n_0}{n}$, respectively. One can also directly convert the test statistic for the independent samples t-test to the point-biserial correlation coefficient with

$$r_{pb} = \frac{t}{\sqrt{t^2 + m}}. \tag{7}$$

For the example data, Equation [5] yields a point-biserial correlation coefficient of

$$r_{pb} = \frac{0.75}{\sqrt{0.75^2 + 3.6}} = 0.37.$$

This value is identical to

$$r_{pb} = \left(\frac{4.7 - 3.6}{1.531}\right)\sqrt{\frac{20(0.5)(0.5)}{20 - 1}} = 0.37,$$

using Equation [6] and the fact that $s_y = 1.531$. Similarly, Equation [7] yields the identical value of

$$r_{pb} = \frac{1.68}{\sqrt{1.68^2 + 18}} = 0.37.$$

One can also easily confirm that the exact same value is obtained when directly computing the product–moment correlation coefficient based on variables x_b and y (i.e., $r_{pb} = cor(x_b, y) = 0.37$) using Equation [2].[1]

It needs to be emphasized here that r_{pb} does not estimate the correlation between the continuous variables X and Y. In other words, the point-biserial correlation coefficient is not an estimate of ρ and therefore cannot be directly compared with a product–moment correlation coefficient that is computed based on a non-dichotomized x and the corresponding y variable. Consequently, combining r and r_{pb} values in a single meta-analysis would not be appropriate.

To obtain direct commensurability, the point-biserial correlation must first be transformed into the *biserial correlation coefficient*, which can be accomplished with

$$r_b = \frac{\sqrt{pq}}{f(z_p)} r_{pb}, \tag{8}$$

where $f(z_p)$ denotes the density of the standard normal distribution at value z_p, which is the point for which $P(Z > z_p) = p$, with Z denoting a random variable following a standard normal distribution (see Figure 1 for an example where $p = 0.3$). The point-biserial correlation coefficient is an estimate of the product-moment correlation coefficient ρ and can be logically compared with estimates of ρ obtained directly from a non-dichotomized x variable and the corresponding y variable.

Continuing with our example, recall that $r_{pb} = 0.37$. In this example, $p = q = 0.5$, implying $z_p = 0$ (i.e., 50% of the area under a standard normal distribution falls above a z-score of 0), and consequently, $f(z_p) = 1/\sqrt{2\pi} \approx 0.399$. The value of r_b can then be computed using Equation [8], yielding

[1]The value of h in equation [5] is often approximated in the literature by $h \approx \frac{n}{n_1} + \frac{n}{n_0}$, which would be equal to 4 in this example and then yields a value of $r_{pb} = 0.35$. However, the exact value of h can be easily computed and is preferable, as it yields the actual value of the point-biserial correlation coefficient and not just an approximation.

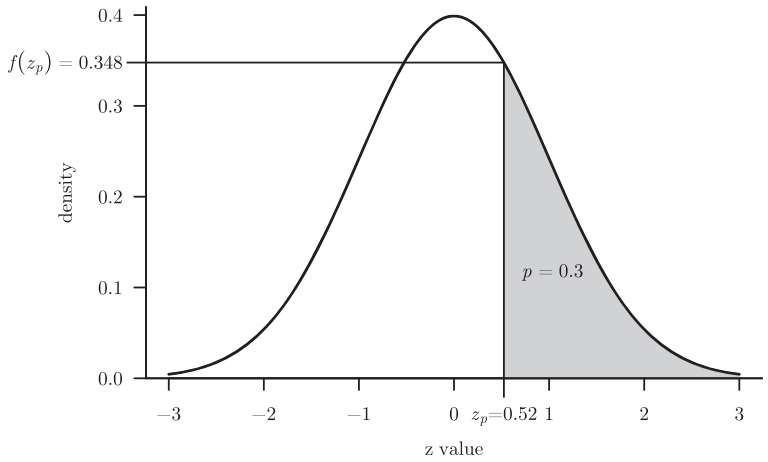

Figure 1. Density of a standard normal distribution. The gray-shaded area corresponds to an area of $p = 0.30$ in the upper tail. The corresponding z-value is $z_p = 0.52$. The density at this z-value is equal to $f(z_p) = 0.348$.

$$r_b = \frac{\sqrt{(0.5)(0.5)}}{0.399} 0.37 = 0.46.$$

There is little difference between $r_b = 0.46$ and $r = 0.44$ in this example, and each value provides an estimate of $\rho = 0.5$. However, in practice, only r_b would be available for inclusion in a meta-analysis if authors dichotomized x and reported stratified summary statistics as described earlier.

3. Properties of the biserial correlation coefficient

Several properties can be noted about r_b. First of all, assuming that X and Y follow a bivariate normal distribution, r_b is a consistent and asymptotically normally distributed estimator of ρ (Soper, 1914; Tate, 1955b). Moreover, based on derivations by Soper (1914), r_b should be unbiased for $\rho = 0$, but bias of order $1/n$ is expected as $|\rho|$ increases. Generally, the bias will be positive (away from 0) but can also be negative (towards 0) when the dichotomization results in more unbalanced group sizes (i.e., as p either approaches 0 or 1). However, in practice, the bias should typically be negligible (Soper, 1914).

The multiplicative factor in front of r_{pb} in Equation [8] is always larger than 1, so that $|r_b| > |r_{pb}|$. Therefore, if $|r_{pb}|$ is large to begin with, it can happen that $|r_b|$ is larger than 1 (Tate, 1955b). This is more likely to happen when p is either close to 0 or 1 (i.e., when the group sizes are very unbalanced). Such a finding may be somewhat unsettling, given that we know that ρ must lie somewhere between -1 and 1. Consequently, when reporting the results from an individual study, one could consider constraining the value of r_b to 1 or -1, depending on the sign of r_b. However, doing so would introduce some degree of negative bias into r_b, although this should again be negligible.

It is also worth noting that r_b described earlier is not the maximum likelihood estimator of ρ (Tate, 1955b). Computation of the maximum likelihood estimator requires iterative methods and, more importantly, access to the raw data for x_b and y. Because raw data can be difficult to obtain for secondary analyses (Wicherts *et al.*, 2006), r_b is often the only option for estimating ρ because it only requires access to the stratified summary statistics (or the t-statistic), which are likely to be reported in research articles.[2]

In the example earlier, dichotomization was carried out by means of a method sometimes referred to as a "median split," in which the sample is ordered along variable x and then divided at the middlemost observation into two groups of equal size (or approximately equal size when n is odd-numbered). This approach is in fact often used for dichotomization, for example, in many branches of psychological research. We will refer to this practice as dichotomization at an *adaptive cutoff*, as the cutoff point along the scale of X is not fixed a priori but depends on the sample characteristics, in particular the value of the sample median. Note that the median is only one example of a value that can be used as an adaptive cutoff. Although less common in applied research, cutoffs may also be set at other percentiles of the sample distribution. For example, a cutoff at the 25th percentile would allocate the lower 25% of the sample to one group and the upper 75% to the other.

In contrast to an adaptive cutoff, a *hard cutoff* may be used and is more common in medical research. Here, the cutoff point c is fixed before data collection (which often represents some commonly agreed upon value of clinical and/or practical relevance) and those observations falling below this predefined cutoff are allocated to one and

[2]If study authors have actually measured x (i.e., non-dichotomously) before turning it into x_b for the analyses, one could also ask the authors for the value of the product–moment correlation r directly, instead of asking for the raw data.

those falling above to the other group. Consider, for example, the Beck Depression Inventory-II (Beck *et al*., 1996), a diagnostic scale to detect depression that is widely used across clinical psychology. The scale yields a score that ranges from 0 to 63, with higher scores indicating more severe levels of depression. Patients scoring 19 or below on this scale are typically classified as *minimally* or *mildly depressed*, whereas those scoring 20 and above are classified as *moderately* or *severely depressed*, so in essence, $c = 19.5$ would constitute the hard cutoff value when adopting this guideline.

From a statistical point of view, the essential difference between the two approaches lies in how n_1 and n_0 (and hence, h, p, q, z_p, and $f(z_p)$) should be treated. For a given total sample size n, an adaptive cutoff at a certain percentile implies that each group's sample size (i.e., n_1 and n_0) is a fixed constant (and, hence, so are the values calculated based on the group sizes). For a hard cutoff, however, n_1 and n_0 are random variables because the number of individuals falling into each group is not predetermined. We will return to the relevance of this point in the discussion.

4. Sampling distribution of the biserial correlation coefficient

Standard meta-analytic procedures (e.g., Shadish and Haddock, 2009) are based on two important assumptions, namely, that the sampling distributions underlying the estimates are normal and that the sampling variances of the estimates are known constants. In practice, these assumptions are hardly ever fully satisfied. For most effect size or outcome measures used in meta-analyses (e.g., standardized mean differences, correlation coefficients, and log odds ratios), we must rely on the asymptotic behavior of the effect size or outcome measure to assume approximately normal sampling distributions and to obtain estimates of the sampling variances that can be treated as known constants for practical purposes.

Consider first the product–moment correlation coefficient r. Assuming that X and Y follow a bivariate normal distribution, the asymptotic sampling distribution of r is in fact normal with expected value ρ and variance

$$Var(r) \overset{\infty}{=} \frac{(1-\rho^2)^2}{n} \tag{9}$$

(e.g., Hedges and Olkin, 1985). In practice, an estimate of the sampling variance of r is typically obtained by replacing ρ with the observed value of r and using $n-1$ in place of n (e.g., Borenstein, 2009), yielding

$$Var(r) = \frac{(1-r^2)^2}{n-1}. \tag{10}$$

Returning to our example from the previous section, a value of $r = 0.44$ was found. We therefore find an estimate for the sampling variance of r equal to $Var(r) = \left(1-0.44^2\right)^2/(20-1) = 0.034$.

Turning now to the biserial correlation coefficient and again assuming a bivariate normal distribution for X and Y, it can again be shown that r_b has a normal sampling distribution asymptotically with expected value ρ and variance

$$Var(r_b) \overset{\infty}{=} \frac{1}{n}\left(\rho^4 + \rho^2\left(\frac{PQz_P^2}{f(z_P)^2} + \frac{(P-Q)z_P}{f(z_P)} - \frac{5}{2}\right) + \frac{PQ}{f(z_P)^2}\right), \tag{11}$$

where P, Q, z_P, and $f(z_P)$ are the population counterparts to variables p, q, z_p, and $f(z_p)$ as defined earlier (Soper, 1914; Tate, 1955b). An estimate of the sampling variance can be obtained with

$$Var(r_b)^{sop} = \frac{1}{n-1}\left(r_b^4 + r_b^2\left(\frac{pqz_p^2}{f(z_p)^2} + \frac{(p-q)z_p}{f(z_p)} - \frac{5}{2}\right) + \frac{pq}{f(z_p)^2}\right). \tag{12}$$

Analogous to Equation [10], we have replaced n by $n-1$ in Equation [12] as this appears to yield a more accurate estimate of the sampling variance. We will return to this issue in the discussion.

We shall refer to Equation [12] as *Soper's exact method*. Using this method, the sampling variance of r_b in the example earlier is estimated to be

$$Var(r_b)^{sop} = \frac{1}{20-1}\left(0.46^4 + 0.46^2\left(0 + 0 - \frac{5}{2}\right) + \frac{(0.5)(0.5)}{0.399^2}\right) = 0.057,$$

where the first two terms within the inner parentheses are equal to zero because $z_p = 0$. Note that the estimated variance of the biserial correlation is larger than that of the product moment correlation (by a factor of $0.057/0.034 \approx 1.68$, that is, 68% larger), which reflects the loss of precision in estimating ρ because of the dichotomization of one of the two variables.

When $|r_b| > 1$, it can happen that Equation [12] yields a negative value, which falls outside of the parameter space for a variance. Because Equation [12] is obtained by substituting r_b for ρ in the equation of the asymptotic

sampling variance of r_b (cf. Tate, 1955b) and $|\rho| \leq 1$, we therefore suggest truncating values of $|r_b| > 1$ to 1 before computing the sampling variance of r_b. Equation [12] will then always yield a positive estimate of the sampling variance, even when $|r_b| > 1$, which properly reflects the fact that there is still some imprecision left in the estimate of ρ, even in such an extreme situation.

To facilitate calculations, Soper (1914) proposed an approximation to Equation [12], given by

$$Var(r_b)^{app} = \frac{1}{n-1}\left(\frac{\sqrt{pq}}{f(z_p)} - r_b^2\right)^2, \tag{13}$$

which we shall refer to as *Soper's approximate method*. Using this approach, the sampling variance of r_b in the example earlier is estimated to be

$$Var(r_b)^{app} = \frac{1}{20-1}\left(\frac{\sqrt{(0.5)(0.5)}}{0.399} - 0.46^2\right)^2 = 0.057,$$

which happens to be identical to $Var(r_b)^{sop} = 0.057$ (when rounded to three digits) in the present case. Again, we suggest truncating values of $|r_b| > 1$ to 1 before using Equation [13].

An alternative method for estimating the sampling variance of r_b was described by Hunter and Schmidt (1990, 2004), who propose to estimate the sampling variance with

$$Var(r_b)^{hs} = \left(\frac{pq}{f(z_p)^2}\right)\frac{\left(1 - r_{pb}^2\right)^2}{n-1}, \tag{14}$$

which we will refer to as the *Hunter and Schmidt method*. The method essentially uses Equation [10] to estimate the variance of the point-biserial correlation coefficient and then corrects this estimate by the square of the factor that converts the point-biserial into the biserial correlation (cf. Equation [8]). This method, applied to the example earlier, yields an estimate of

$$Var(r_b)^{hs} = \left(\frac{(0.5)(0.5)}{0.399^2}\right)\frac{\left(1 - 0.37^2\right)^2}{20-1} = 0.062,$$

slightly larger than the values found based on $Var(r_b)^{sop}$ and $Var(r_b)^{app}$.

Finally, some researchers may be inclined to use the equation for the sampling variance of the product–moment correlation coefficient (i.e., Equation [10]) for calculating the sampling variance of the biserial correlation coefficient. This is also what would happen if a biserial correlation is entered into meta-analytic software that then treats this value as if it were a regular product–moment correlation coefficient. Thus, the sampling variance of r_b would then be estimated with

$$Var(r_b)^{r} = \frac{\left(1 - r_b^2\right)^2}{n-1}. \tag{15}$$

We shall refer to this approach as the *naive method*, yielding an estimate of

$$Var(r_b)^{r} = \frac{\left(1 - 0.46^2\right)^2}{20-1} = 0.033$$

when applied to the example earlier, a value that is substantially smaller than any of the estimates obtained earlier.

5. Inference for ρ based on biserial data

A test of the null hypothesis H_0: $\rho = \rho_0$ could in principle be obtained by computing the Wald-type test statistic $z = (r_b - \rho_0)/SE(r_b)$, where $SE(r_b) = \sqrt{Var(r_b)}$ is estimated using any of the methods described earlier. The resulting value of z can then be compared against the critical values of a standard normal distribution (e.g., ± 1.96 for $\alpha = 0.05$, two-sided). When $Var(r_b)$ is a consistent estimator of the sampling variance of r_b, such a test should perform nominally asymptotically but may not adequately control the Type I error rate in small samples.[3]

[3]Better control of the Type I error rate should be possible by substituting ρ_0 for r_b in the equations for $Var(r_b)$, so that the variance, and hence the standard error, is estimated under H_0 and therefore involves one less unknown quantity.

However, a hypothesis test only provides a dichotomous decision about H_0: $\rho = \rho_0$. More interesting is the question of how to obtain a confidence interval for ρ based on the biserial correlation coefficient. Applying the same principle used to compute the test statistic earlier, the bounds of a 95% confidence interval could be obtained with

$$r_b \pm 1.96\, SE(r_b), \tag{16}$$

where $SE(r_b)$ is again estimated using any of the methods described in the previous section. Asymptotically, such a confidence interval should have nominal coverage probability if $Var(r_b)$ is a consistent estimator of the sampling variance of r_b.

Returning to the example and using Soper's equations for estimating the sampling variance, we would obtain the bounds $0.46 \pm 1.96\sqrt{0.057} = (-0.01, 0.93)$ for a 95% confidence interval for ρ. On the other hand, using the approach by Hunter and Schmidt, we would obtain $0.46 \pm 1.96\sqrt{0.062} = (-0.03, 0.95)$. Finally, the naive method would yield the bounds $0.46 \pm 1.96\sqrt{0.033} = (0.10, 0.82)$.

Alternatively, we suggest the use of a variance-stabilizing transformation analogous to the *r*-to-*z* transformation for product–moment correlations.[4] Starting with Equation [13] and applying the delta method, it can be shown that

$$z_{r_b} = g(r_b) = \left(\frac{a}{2}\right) ln\left(\frac{1 + ar_b}{1 - ar_b}\right) \tag{17}$$

follows asymptotically a normal distribution with expected value $g(\rho)$ and variance $1/(n-1)$, where $a = \sqrt{f(z_p)}/\sqrt[4]{pq}$. Then, 95% confidence interval bounds for $g(\rho)$ can be computed with

$$z_{r_b} \pm 1.96\sqrt{1/(n-1)}, \tag{18}$$

which can be back-transformed to bounds for ρ by means of the inverse transformation

$$r_b = g^{-1}(z_{r_b}) = \left(\frac{1}{a}\right)\frac{exp(2z_{r_b}/a) - 1}{exp(2z_{r_b}/a) + 1}. \tag{19}$$

To avoid problems in computing [17] when $|r_b| > 1$, we again suggest truncating such values to ± 1.

In the example, $a = \sqrt{0.399}/\sqrt[4]{(0.5)(0.5)} = 0.893$, and the value of the transformed biserial correlation coefficient is therefore

$$z_{r_b} = \left(\frac{0.893}{2}\right) ln\left(\frac{1 + 0.893(0.46)}{1 - 0.893(0.46)}\right) = 0.39.$$

We then obtain the bounds $0.39 \pm 1.96\sqrt{1/(20-1)} = (-0.06, 0.84)$ in the transformed metric. Finally, after applying the inverse transformation function [19] to these bounds, we obtain a 95% confidence interval for ρ with bounds $(-0.08, 0.82)$. Note that this and all of the confidence intervals obtained earlier are very wide (in part because of the small sample size of the example), reflecting large uncertainty about the value of ρ.

6. Monte Carlo simulation study

As described earlier, several different methods are available for estimating the sampling variance of the biserial correlation coefficient. All of these methods are based on various simplifications and large-sample approximations, and their accuracy in practice is largely unknown. A search of the literature revealed only two papers that have examined the accuracy of $Var(r_b)^{sop}$ (Equation [12]) for practical purposes. Lord (1963) conducted a small simulation study with $n = 100$, $\rho \in \{0.5, 0.9\}$, two different cutoff values (corresponding to $P \in \{0.10, 0.38\}$), and 500 iterations per condition. The simulation study by Koopman (1983) based on $n = 400$, $\rho \in \{0, 0.05, 0.10, \ldots, 0.95\}$, four different cutoff values (corresponding to $P \in \{0.05, 0.10, 0.25, 0.50\}$), and 20,000 iterations was more comprehensive. Close agreement between the actual variance of r_b and the estimated sampling variance based on Equation [12] was found, except when the groups become very unbalanced (i.e., $P \leq 0.10$) and ρ is very large ($|\rho| \geq 0.90$).

Results for smaller or larger values of n and for the other methods are currently not available. Moreover, the derivation of the asymptotic sampling variance of the biserial correlation coefficient by Soper (1914) and the simulation studies by Lord (1963) and Koopman (1983) were based on the use of hard cutoff values. It is unclear

[4] A simpler variance-stabilizing transformation for biserial correlations was derived by Tate (1955a), which is only applicable when $p = 0.5$. A derivation of the transformation suggested here is given in the Supporting Information.

whether Equations [12] and [13] provide appropriate estimates of the sampling variance when an adaptive cutoff is used. Finally, the performance of the various methods for constructing confidence intervals for ρ based on the biserial correlation coefficient has not been investigated to our knowledge.

We expect the naive method (i.e., Equation [15]) to underestimate the actual variance, as it treats biserial correlations as if they were product–moment correlations and therefore ignores the loss of precision resulting from the dichotomization. However, the extent of the underestimation is unclear. Also, the appropriateness of the Hunter and Schmidt method has, to our knowledge, not been examined in prior work. We therefore conducted a simulation study to examine the accuracy of the various approaches for estimating the sampling variance of a biserial correlation coefficient and the performance of the various methods for constructing a corresponding confidence interval.

The simulation study was programmed in R (R Core Team, 2014) and used a full factorial design with

- 6 different values for the population correlation: $\rho \in \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9\}$,
- 12 different values for the sample size: $n \in \{20, 30, 40, 60, 80, 100, 150, 200, 250, 300, 600, 1200\}$, and
- 5 different cutoff proportions: $P \in \{0.1, 0.2, 0.5, 0.8, 0.9\}$,

resulting in a total of 360 different conditions. For each condition, 100,000 iterations were carried out. Even though very large correlation coefficients and/or sample sizes are likely to be rare in practice (e.g., Bosco *et al*., 2015; Meyer *et al*., 2001; Richard *et al*., 2003), our goal was to provide a comprehensive examination of the statistical properties of the biserial correlation coefficient under a wide range of conditions. Moreover, including also large values of n in the simulation provides information about the asymptotic behavior of the coefficient.

The simulation proceeded in three stages. First, for a particular value of ρ, we used the function mvnorm() of the MASS package (Venables and Ripley, 2002) to simulate n observed values of x and y from a bivariate normal distribution (with $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$, without loss of generality). At the second stage, variable x was dichotomized — once using an adaptive cutoff value c_A and once using a hard cutoff c_H, yielding the dichotomized variables x_A and x_H, respectively. Specifically, c_A was calculated based on the sample such that a predefined proportion P of observations in the *sample distribution* fell below the cutoff point, and c_H was chosen such that a proportion P of observations in the *population distribution* would fall below the cutoff point. Therefore, for $P = 0.5$, the adaptive cutoff value c_A was the sample median (i.e., 50th sample percentile), while the hard cutoff value c_H was 0 (i.e., the population median of the random variable X). For $P = 0.2$ and $P = 0.8$, c_A was set equal to the 20th and 80th sample percentile, respectively, while c_H was set to -0.84 and 0.84, respectively. Finally, for $P = 0.1$ and $P = 0.9$, the 10th and 90th sample percentiles were used for c_A and values of -1.28 and 1.28 for c_H.

Note that the actual proportion of observations in the *sample* falling below c_H may differ from P in any particular iteration because of chance. As P is moved further away from 0.5 and n is small, it becomes increasingly more likely that the group sizes become so unbalanced that n_1 or n_0 could be equal to 0 or 1 when using a hard cutoff. The case where one of the groups is empty is degenerate because there is not even a way of estimating r_b in such a situation. When one of the groups consists of a single observation, one could in principle apply the methods, but preliminarily simulations indicated that the methods simply break down in such an extreme case. Therefore, in each iteration, the number of observations within each level of the dichotomized variable (i.e., the sample sizes of the two groups) was checked. If either n_1 or n_0 was less than 2 when using a hard cutoff, the iteration was discarded, and an additional iteration was run.[5] We return to the implications of this decision in the discussion section.

At the final stage, we computed the standardized mean difference of variable y based on the two groups formed by the levels of variable x_A using Equation [3]. This standardized mean difference was then converted into the point-biserial correlation r_{pb} using Equation [5], which was then further converted into the biserial correlation using Equation [8]. We denote the point-biserial correlation and the biserial correlation based on the adaptive cutoff as r_{pb}^A and r_b^A, respectively. The same procedure was then applied based on the dichotomized variable x_H, which yielded r_{pb}^H and r_b^H.

The sampling variances of r_b^A and r_b^H were then estimated in four different ways. First, Soper's exact method (i.e., Equation [12]) was used, producing $Var\left(r_b^A\right)^{sop}$ and $Var\left(r_b^H\right)^{sop}$. Second, Soper's approximate method (i.e., Equation [13]) was used to compute $Var\left(r_b^A\right)^{app}$ and $Var\left(r_b^H\right)^{app}$. Third, the Hunter and Schmidt method (i.e., Equation [14]) was used, yielding $Var\left(r_b^A\right)^{hs}$ and $Var\left(r_b^H\right)^{hs}$. Finally, the naive method (i.e., Equation [15]) was used to compute $Var\left(r_b^A\right)^{r}$ and $Var\left(r_b^H\right)^{r}$.

Using r_b^A and the four different estimates of $Var\left(r_b^A\right)$, confidence intervals for ρ (i.e., based on Equation [16]). The resulting confidence intervals are denoted $CI\left(r_b^A\right)^{sop}$, $CI\left(r_b^A\right)^{app}$, $CI\left(r_b^A\right)^{hs}$, $CI\left(r_b^A\right)^{r}$ for Soper's exact and approximate

[5]The probability of this occurring can be easily calculated based on the binomial distribution with parameters n and P. For example, for $n = 20$, the probability of observing $n_0 \in \{0, 1, n-1, n\}$ is less than 0.01% for $P = 0.5$, about 7% for $P \in \{0.2, 0.8\}$, and about 39% for $P \in \{0.1, 0.9\}$. For $n \geq 80$, all probabilities are below 1% and essentially negligible.

methods, the method suggested by Hunter and Schmidt, and the naive method, respectively. In addition, a fifth confidence interval, $CI\left(r_b^A\right)^{vs}$, was computed using the variance-stabilizing transformation as described by Equations [17], [18], and [19]. Confidence intervals based on r_b^H were obtained analogously.

The entire procedure was repeated for each of the 100,000 iterations in each of the 360 conditions. After the 100,000 iterations within a particular condition were completed, we recorded the means of the point-biserial and biserial correlation coefficients (denoted by $\bar{r}_{pb}^A$, $\bar{r}_b^A$, $\bar{r}_{pb}^H$, and $\bar{r}_b^H$) and the observed variances of r_b^A and r_b^H (denoted by $s^2_{r_b^A}$ and $s^2_{r_b^H}$). We also computed the mean of the 100,000 sampling variance estimates for each of the estimators described earlier (denoted, for example, by $\overline{Var}\left(r_b^A\right)^{sop}$ and the other values analogously). Finally, for each method of constructing a confidence interval, we recorded in what percentage of iterations the computed confidence interval contained the true value of ρ (i.e., the empirical coverage probability).

We then examined the absolute and relative bias in the r_{pb}^A, r_b^A, r_{pb}^H, and r_b^H values (e.g., $\text{Bias}\left(r_b^A\right) = \bar{r}_b^A - \rho$ and Relative $\text{Bias}\left(r_b^A\right) = \left(\bar{r}_b^A - \rho\right)/\rho$). In addition, to examine how well the various variance estimators approximate the actual variance of the r_b^A and r_b^H values within a particular condition, we computed the ratio of the mean sampling variance to the actually observed variance for a particular method (e.g., $\overline{Var}\left(r_b^A\right)^{sop}/s^2_{r_b^A}$). A ratio equal to 1 indicates that a particular method provides an unbiased estimate of the actual variance. Values above 1 indicate that a method overestimates the actual sampling variance, while values below 1 indicate that the actual sampling variance is underestimated on average. Finally, to assess the quality of the various methods for constructing confidence intervals, we compared the empirical coverage probabilities of the various confidence intervals to the nominal 95% rate. Confidence intervals that actually cover the true ρ in 95% of the cases show nominal performance, whereas values above or below 95% indicate overcoverage or undercoverage, respectively.

7. Results

The results of the simulation study are displayed in Figures 2–5, and their description is structured in three parts. The first part focuses on the bias of the estimators and illustrates the crucial distinction between the point-biserial and the biserial correlation coefficient. In the second part, we assess the quality of the different methods for estimating the sampling variance of the biserial correlation coefficient. Finally, the third part pertains to the coverage accuracy of the different confidence intervals. As described earlier, the simulation study was conducted with $P \in \{0.1, 0.2, 0.5, 0.8, 0.9\}$. A comparison of the results for $P = 0.2$ and $P = 0.8$ and for $P = 0.1$ and $P = 0.9$ showed that they were virtually identical. For conciseness, the results are therefore discussed for $P = 0.5$, $P = 0.2$, and $P = 0.1$ only. In addition, an examination of the results showed that a sample size of 300 was typically sufficient to establish the asymptotic behavior of the different estimators/methods. We therefore only show results for values up to $n = 300$.

7.1. Bias of the estimators

Figure 2 shows the bias in the biserial and point-biserial correlation coefficient based on a hard and adaptive cutoff (i.e., the bias of r_b^H, r_b^A, r_{pb}^H, and r_{pb}^A) as estimators for ρ when $P = 0.5$. Corresponding figures for $P = 0.2$ and $P = 0.1$ are given as part of the Supporting Information as Figures S1 and S2, respectively. Overall, the biserial correlation coefficient is largely free of bias, irrespective of whether a hard cutoff (black line with circles) or an adaptive cutoff (gray line with circles) was used for dichotomization. Slight bias can be observed for large values of ρ when sample sizes are small: When an adaptive cutoff is used, ρ is slightly underestimated, whereas slight overestimation can usually be observed when a hard cutoff is used for the dichotomization. However, these biases rapidly diminish as the sample size increases. For $n \geq 60$, the bias is essentially negligible, as the biserial correlation coefficient then deviates from ρ by less than 1% on average.

A different picture emerges for the performance of the (incorrectly used) point-biserial correlation coefficient, displayed by the black and gray lines with squares in Figure 2 (and Figures S1 and S2). Except when $\rho = 0$, the point-biserial correlation coefficient underestimates the population correlation markedly. This negative bias increases for larger values of ρ and exists irrespective of how the sample was dichotomized. As ρ increases, so does the absolute bias, albeit at different rates for different cutoff points. For example, ρ is persistently underestimated by approximately 20% to 23% when dichotomization occurs at the median, but this relative bias exacerbates as the cutoff point is moved away from the median: For $P = 0.2$, the relative bias falls approximately between 29% and 32%, and for $P = 0.1$, between 32% and 44%. Importantly, the bias does not diminish as the sample size increases, demonstrating the inconsistency of the point-biserial correlation in estimating the product–moment correlation ρ.

7.2. Variance of the estimators

Figure 3 shows $\overline{Var}\left(r_b^A\right)^r/s^2_{r_b^A}$, that is, the ratio of the mean of the sampling variance estimates using the naive method (i.e., Equation [15]) to the actually observed variance in the biserial correlation coefficients as a function

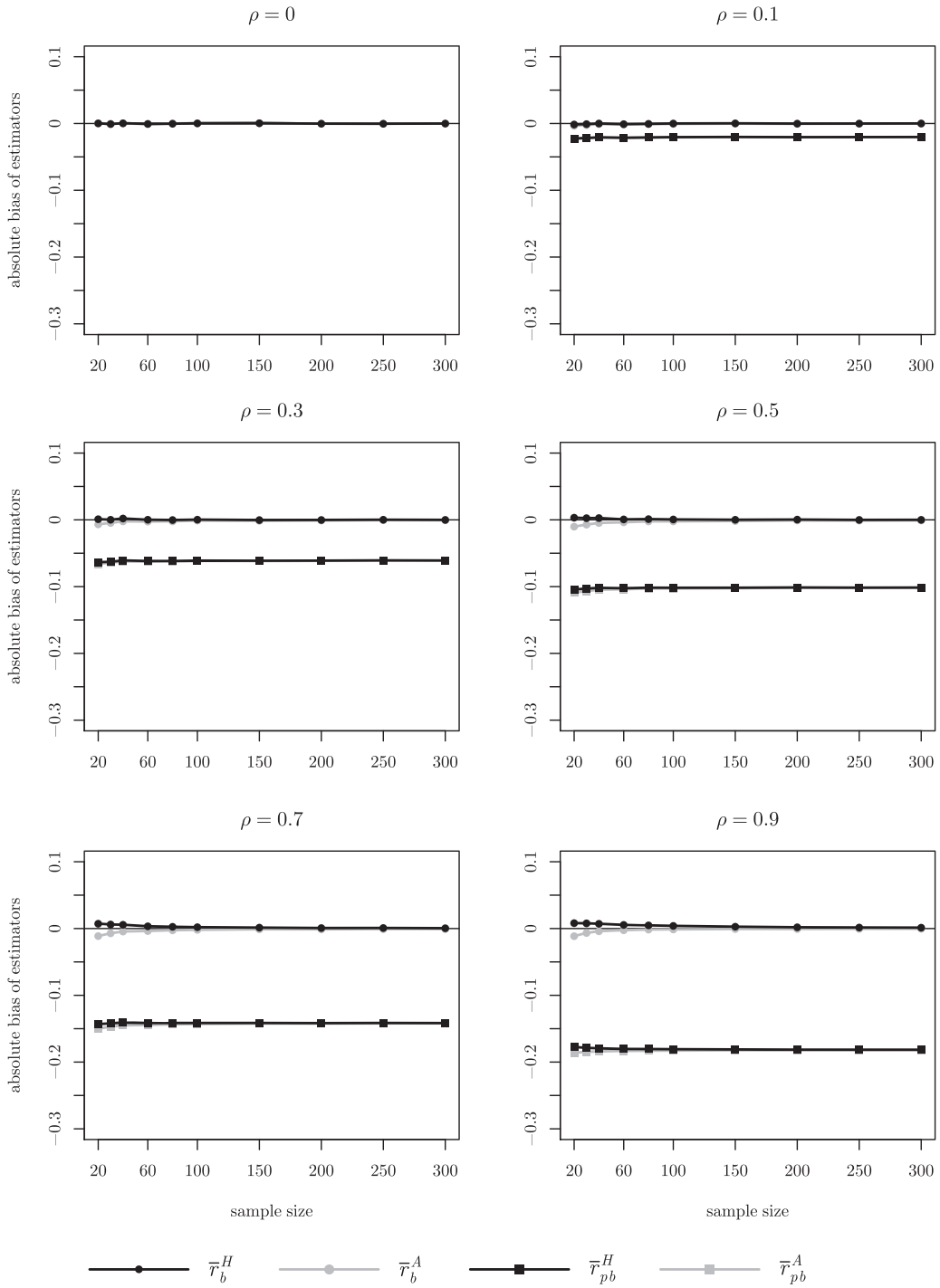

Figure 2. Absolute bias of the biserial and point-biserial correlation as a function of n for different values of ρ with $P = 0.5$.

of n for $\rho = 0$ (black lines) and $\rho = 0.9$ (gray lines) when using an adaptive cutoff. For $\rho = 0$, the true variance is already underestimated by 37% to 46% for $P = 0.5$, by 52% to 60% for $P = 0.2$, and by 67% to 74% for $P = 0.1$. As ρ increases, the underestimation becomes even more severe, eventually growing to 94% (for $\rho = 0.9$, $n = 300$, and $P = 0.1$; see the gray line with triangles). This poor performance is comparable when a hard instead of an adaptive cutoff is used and shows no considerable improvement as the sample size increases.

Leaving aside the naive estimator of the sampling variance, Figure 4 shows the performance of the other estimators for $P = 0.5$ (Figures S3 and S4 show the results for $P = 0.2$ and $P = 0.1$, respectively). For $\rho = 0$, the figures show that all methods tend to underestimate the actual sampling variance of the biserial correlation at low sample sizes. However, all estimation methods produce consistent estimates that converge to the actual sampling variance as the sample size increases. Thus, for $n \geq 100$, all of the methods produce estimates that deviate on average by no more than four percent from the actual sampling variance when $\rho = 0$.

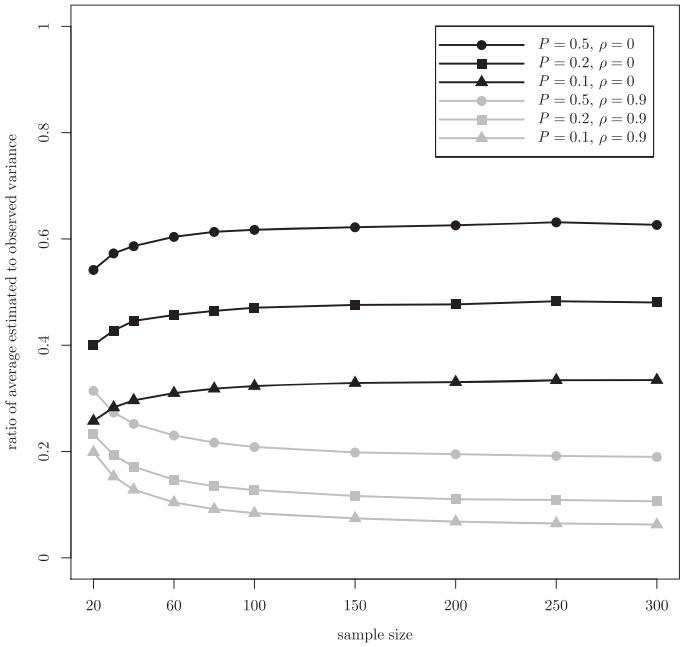

Figure 3. Ratio of the average estimated variance based on the naive method to the actually observed variance (i.e., $\overline{Var}\left(r_b^A\right)^r / s^2_{r_b^A}$) as a function of n for different values of ρ and P.

Beyond $\rho = 0$, however, only the two estimation methods suggested by Soper appear to provide consistent estimates across most conditions. In particular, neither the exact nor the approximate method (lines with circles and squares, respectively) yields variance estimates that deviate by more than 5% from the actual variance for $n \geq 80$, the only exception being conditions with $\rho = 0.9$ and $P = 0.1$ (Figure S4).

Both methods typically underestimate the true sampling variance at smaller sample sizes, but some overestimation occurs in a few cases, especially when ρ is very large. In fact, the largest deviation is found for the extreme situation of $n = 20$, $\rho = 0.9$, and $P = 0.1$, where the estimated sampling variance of biserial correlations based on an adaptive cutoff is, on average, 1.52 times larger than its actual value when using the approximate method. Overall, discernible differences between the exact and the approximate method are only found when ρ is large and n is small, whereas they are negligible in most other conditions.

In contrast, the method suggested by Hunter and Schmidt yields consistent estimates only for very small population correlations (lines with triangles in Figures 4, S3, S4). For $\rho \geq 0.3$, however, the method increasingly begins to produce values that overestimate the sampling variance and do not converge to the actual variance as the sample size increases. The degree of overestimation increases with ρ and can grow alarmingly high, yielding estimates that are, on average, more than twice as large as the actual variance (Figure S4).

7.3. Coverage probability of the confidence intervals

For biserial correlation coefficients computed based on an adaptive cutoff, Figure 5 displays the empirical coverage probabilities of the various methods for constructing confidence intervals described earlier when $P = 0.5$ (results for $P = 0.2$ and $P = 0.1$ are shown in Figures S5 and S6, respectively). Results based on the naive method have been left out because of the extremely poor performance. Results for confidence intervals based on a hard cutoff were similar and hence are not shown.

For the two methods based on Soper (black and gray lines with circles) and the variance-stabilizing transformation (black line with squares), we observe convergence to the nominal coverage level of 95% as the sample size increases, irrespective of the population correlation (note that the lines for the confidence intervals based on Soper's exact and approximate method are often indistinguishable). The speed of convergence decreases slightly as ρ rises, although for $n \geq 100$, all three methods achieve coverage probabilities that deviate from 95% by less than two percentage points in all conditions. For $n \geq 200$, all deviations are less than one percentage point.

When sample sizes are smaller, all three methods yield confidence intervals that usually undercover, except for the confidence intervals based on the variance-stabilizing transformation, which have higher than nominal coverage when $\rho = 0.9$ and sample sizes are small. Considering all conditions together, the variance-stabilizing method yields confidence intervals with the closest to nominal performance in the largest number of cases.

The errors of the Hunter and Schmidt and the naive methods in estimating the sampling variance result in analogous deficiencies in the confidence intervals based on these methods. In particular, except when ρ is low,

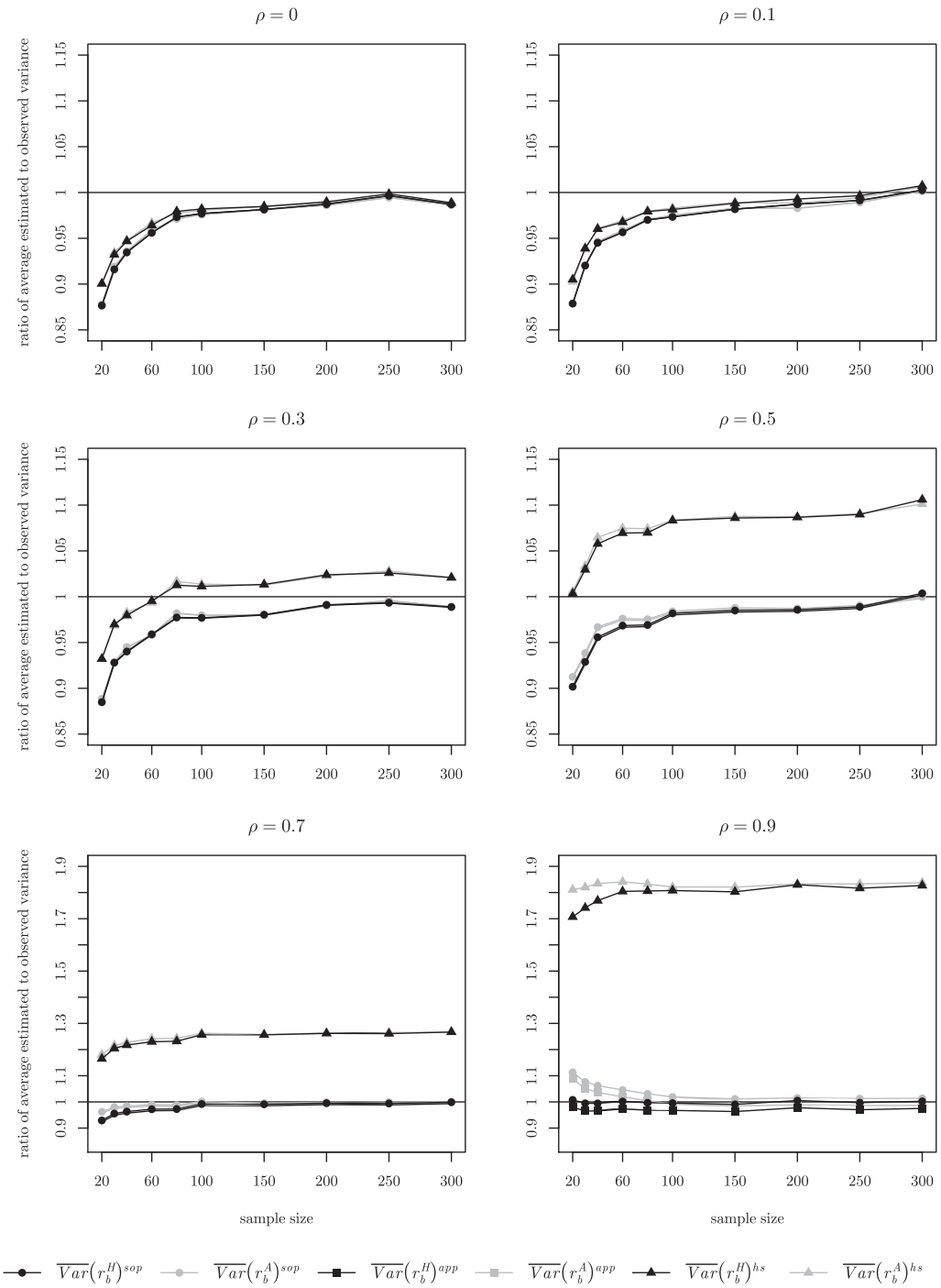

Figure 4. Ratio of the average estimated to observed variance as a function of n for different values of ρ with $P = 0.5$.

the Hunter and Schmidt method typically overestimates the sampling variance and in turn tends to produce confidence intervals that are too wide on average (black lines with triangles in Figures 5, S5, S6). Moreover, for $\rho \geq 0.5$, the method produces confidence intervals whose coverage rate does not appear to converge to the nominal confidence level. In contrast, the naive method underestimates the sampling variance, and the resulting confidence intervals are too narrow and undercover in all conditions (with coverage probabilities that never exceeded 88% and that were as low as 34% in some conditions).

8. Discussion

Dichotomization of continuous variables is a common occurrence in empirical research and poses a challenge to meta-analysts wishing to integrate measures of association based on artificially dichotomized data with those

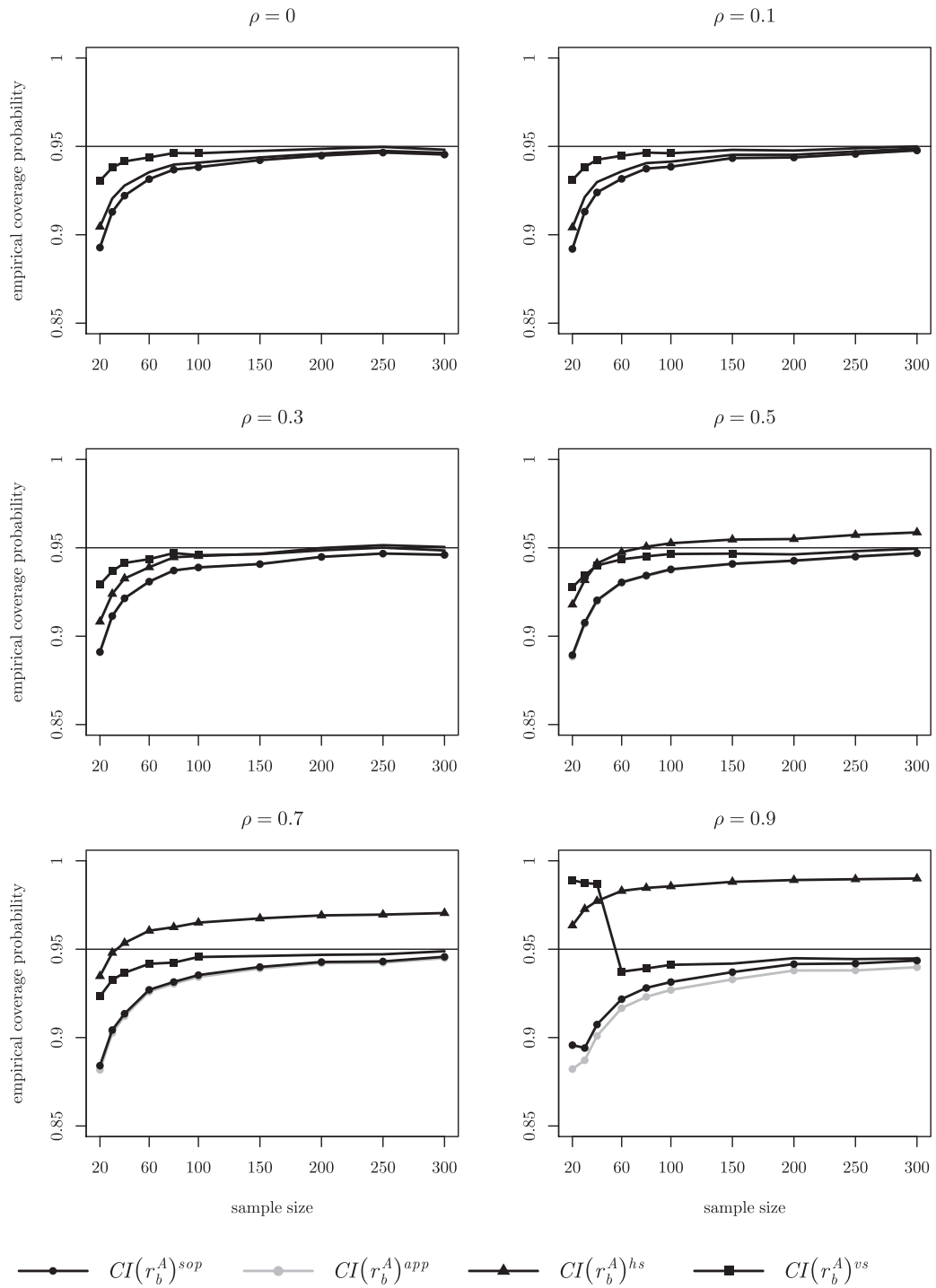

Figure 5. Empirical coverage probability of 95% confidence intervals for different levels of ρ with $P = 0.5$ and dichotomization at adaptive cutoff.

based on continuous data. Meeting this challenge requires the meta-analyst to select an appropriate (1) estimator, (2) measure of the estimator's sampling variance, (3) approach for drawing inferences about individual studies, and (4) approach for integrating the results of the entire collection of studies. In this section, we use the results of the simulation study to examine how each of these steps is affected by the presence of studies in which one or both variables of interest have been artificially dichotomized.

8.1. Choice of estimator

In situations in which one of two continuous variables is measured dichotomously while the other is measured continuously, the biserial correlation coefficient can be computed to estimate the strength of the association

between the two underlying continuous variables. Unlike the point-biserial correlation coefficient, the biserial correlation can be logically compared with product–moment correlation coefficients obtained from data of two variables that were measured on a continuum.

In the present study, we have illustrated this crucial distinction between the biserial and the point-biserial correlations by means of a Monte Carlo simulation. As expected, the results demonstrate a strong negative bias of the point-biserial correlation in estimating the underlying population correlation, whereas the biserial correlation coefficient was largely free of bias. Therefore, the biserial correlation coefficient should be used when correlations based on one continuous variable and one artificially dichotomized variable are to be integrated with correlations based on two continuous variables. In contrast, the point-biserial correlation coefficient should only be used when the dichotomous variable represents a natural/true dichotomy and not an underlying continuous construct.[6]

Similarly, in case both variables are measured dichotomously and the resulting 2×2 table of cell frequencies (or the corresponding odds ratio) is reported, the tetrachoric correlation coefficient can be computed (or approximated), which estimates the association between the underlying continuous variables (Pearson, 1900, 1913), whereas the phi coefficient is only appropriate when both dichotomous variables represent natural/true dichotomies. Therefore, only tetrachoric, biserial, and product–moment correlation coefficients can be logically compared with one another.

8.2. Sampling variance

In order to incorporate a biserial correlation coefficient into a meta-analysis based on standard meta-analytic methods (e.g., Shadish and Haddock, 2009), we must be able to estimate the sampling variance of the coefficient. The present simulation study examined the adequacy of several methods for estimating the sampling variance of the biserial correlation coefficient and found that the exact and approximate methods developed by Soper (1914) offered the best overall performance. Already at moderate sample sizes, the estimates provided by both methods deviate on average only slightly from the observed sampling variance and appear to be asymptotically accurate. Interestingly, only very minor differences were found between the exact and the approximate methods, yielding the latter as a feasible alternative for practical use. One exception to this appears to be the case where P is close to 0 (or 1) and ρ is very large, in which case the exact method typically provides better estimates.

In contrast, applying the equation for the sampling variance of the product–moment correlation to biserial correlations has proven inexpedient. Such a naive approach (as we would call it) produces estimates that considerably underestimate the actual sampling variance. Given that an underestimated sampling variance would lead to inflated Type I error rates and overly narrow confidence intervals, we strongly advise against this practice.

In addition, the present study also examined the accuracy of the method suggested by Hunter and Schmidt (2004) for estimating the sampling variance of biserial correlations. The simulation results show that this method is not consistent when $\rho \neq 0$. In particular, it tends to overestimate the sampling variance — in extreme cases by a factor of more than two. To understand this poor performance, we can compare their proposed method with Soper's approximate method, Equation [13], which can be rearranged to

$$Var(r_b)^{app} = \left(\frac{pq}{f(z_p)^2}\right) \frac{\left(1 - \frac{\sqrt{pq}}{f(z_p)} r_{pb}^2\right)^2}{n-1}. \tag{20}$$

Comparing this with Equation [14] shows that the only difference between Soper's approximate method and the method by Hunter and Schmidt is the multiplicative factor $\sqrt{pq}/f(z_p)$ within the second set of parentheses, which is always larger than 1 (except for the degenerate case where $p = 0$ or 1, but in which case, r_{pb} and r_b cannot be computed in the first place). Therefore, the amount subtracted from 1 is always larger in Soper's approximate method, so that $Var(r_b)^{hs}$ will typically be larger than $Var(r_b)^{app}$.[7] Consequently, the method by Hunter and Schmidt usually overestimates the sampling variance, leading to overly conservative results. Therefore, we advise against the use of this method in the vast majority of research situations.

As mentioned earlier, when using Soper's exact and approximate methods for estimating the sampling variance of the biserial correlation, we suggest using $n - 1$ in the denominator of the equations (cf. Equations [12] and [13]). In part, this mirrors how the sampling variance of the product–moment correlation coefficient is

[6]Whether a dichotomous variable can be considered to reflect an underlying continuum may be debatable in certain situations, as demonstrated by the now famous Pearson–Yule debate (Pearson, 1900; Pearson and Heron, 1913; Yule, 1912).

[7]However, when $\left(\sqrt{pq}/f(z_p) + 1\right) r_{pb}^2 > 2$, the amount subtracted from 1 becomes so large that the squared term becomes larger with Soper's approximate method, so that $Var(r_b)^{app} > Var(r_b)^{hs}$. With $p = 0.5$, this would happen if $|r_{pb}| > 0.9421$. With $p = 0.2$ or 0.8, $|r_{pb}| > 0.9074$ would lead to such a case. Finally, for $p = 0.1$ or 0.9, Soper's approximate method would lead to a larger variance estimate if $|r_{pb}| > 0.8592$. However, in practice, one is unlikely to encounter such extreme values of r_{pb}. A more detailed treatment of this issue is given in the Supporting Information.

typically estimated (cf. Equations [9] and [10]). More importantly, as the results show (Figures 4, S3, S4), the equations based on Soper (1914) typically lead to some underestimation of the actual variance, especially when n is small. Using n in the denominator would result in more severe underestimation, or in other words, using $n-1$ in the denominator helps to reduce the underestimation slightly.

On the other hand, for very large values of ρ, Soper's equations usually lead to overestimation of the sampling variance to some degree when n is small. Using $n-1$ exacerbates this overestimation. However, this problem only becomes apparent for $\rho = 0.9$, which makes it less likely to be relevant for practice than the underestimation when ρ is small. These findings also indicate that there is no simple correction that could be used to make Soper's equations more accurate when n is small, as the amount of overestimation/underestimation depends on ρ and also on p.

8.3. *Inference about individual studies*

Forest plots are frequently used in meta-analyses to present the findings of the individual studies (Schild and Voracek, 2013). For each study, the plot shows the observed value of the outcome measure of interest (e.g., correlation coefficient) and the corresponding confidence interval (usually 95%) for the study's true outcome (Lewis and Clarke, 2001).

In the present simulation study, we examined various methods for constructing confidence intervals for ρ based on the biserial correlation coefficient. The results on the coverage probability of confidence intervals further illustrate the consequences of (in)accurately estimating the sampling variance. In particular, we found that Wald-type confidence intervals based on the exact and the approximate methods by Soper (1914) offered more accurate overall performance than those based on the naive method and the method suggested by Hunter and Schmidt (2004).

In addition to those four methods, we derived and examined a variance-stabilizing transformation for the biserial correlation coefficient. Analogous to Fisher's r-to-z transformation for product–moment correlation coefficients (Fisher, 1921), confidence intervals for ρ are then obtained by first transforming the estimated biserial correlation using Equation [17], calculating confidence intervals in the transformed metric using Equation [18], and then back-transforming the bounds using Equation [19]. The simulation study showed that the confidence intervals for ρ based on this method are more accurate (i.e., have closer to nominal coverage) than confidence intervals calculated directly from the biserial correlation and its estimated sampling variance. Thus, we suggest the use of this transformation when making inferences based on biserial correlation coefficients for individual studies.

8.4. *Integrating the findings*

While the variance-stabilizing transformation is useful for inferential purposes when examining individual studies, it generally cannot be used when aggregating multiple coefficients in a meta-analysis. First of all, when integrating coefficients from studies using different types of correlation coefficients, the respective transformations are different, invalidating any direct comparisons of the transformed coefficients. In particular, the variance-stabilizing transformation for the biserial correlation coefficient differs from Fisher's r-to-z transformation for the product–moment correlation, so that transformed product–moment correlations cannot be directly compared with transformed biserial correlations (however, see Pustejovsky, 2014, who suggests an approach for applying Fisher's r-to-z transformation also to biserial correlations).

Even in the unlikely case where we are only integrating a set of biserial correlation coefficients, comparing and aggregating their transformed values would be inappropriate as the transformations depend on the value of a, which is a function of p (cf. Equation [17]), which in turn is likely to differ across studies. Hence, even if the value of ρ is homogeneous across studies, the transformed true correlations (i.e., the $g(\rho)$ values) are then necessarily heterogeneous. Moreover, back-transforming the results from such a meta-analysis into the correlation metric for easier interpretation by means of Equation [19] is problematic, because a value for a (or rather, p) would have to be chosen and there is no obvious way of doing so when p differs across studies.

Consequently, a meta-analysis involving product–moment and biserial correlation coefficients needs to be based on the raw coefficients. Similarly, tetrachoric correlation coefficients could be added to such a mix without difficulties. All of these coefficients are logically comparable and estimate the same underlying parameter. However, it is possible that the true strength of the relationship (i.e., ρ) differs systematically between studies reporting different types of information/coefficients. Hence, when conducting a meta-analysis based on different types of correlation coefficients, we recommend to record the type of coefficient that was obtained from each study. This variable then allows for an examination of systematic differences across different types of correlation coefficients. This examination can take the form of a descriptive inspection or can be integrated within a more extensive moderator analysis with appropriate meta-regression models (e.g., Raudenbush, 2009).

Table 2. Example data for a meta-analysis with studies of different types.

		Continuous data		Group summary statistics						2 × 2 table				Estimates	
No	Study ID	r_i	n_i	$\bar{y}_{1,i}$	$s_{1,i}$	$n_{1,i}$	$\bar{y}_{0,i}$	$s_{0,i}$	$n_{0,i}$	a_i	b_i	c_i	d_i	y_i	v_i
1	Sivertsen *et al.* (2012), HUNT2	0.606	2427											0.606	0.0002
2	Sivertsen *et al.* (2012), HUNT3	0.568	2844											0.568	0.0002
3	Inocente *et al.* (2014)	0.342	81											0.342	0.0098
4	Taylor *et al.* (2005)	0.726	150											0.726	0.0015
5	Taylor *et al.* (2013)	0.627	93											0.627	0.0040
6	Taylor *et al.* (2016)	0.607	814											0.607	0.0005
7	Lancee *et al.* (2013)			9.46	3.73	281	4.91	2.74	198					0.703	0.0012
8	Jansson and Linton (2006)									55	13	61	83	0.576	0.0074

8.5. *Illustrative example*

As an illustration, we collected data for a meta-analysis on the association between anxiety and depression among persons with sleep difficulties. Table 2 provides the information obtained from eight studies examining this association. The first six studies assessed the association directly using product–moment correlations, while studies 7 and 8 dichotomized one or two of the variables, respectively. Hence, for the first six studies, the estimate of the association is simply the reported product–moment correlation, whose sampling variance can be estimated with Equation [10]. These values are given in the first six rows of the columns denoted by y_i and v_i, respectively.

In study 7, Lancee *et al.* (2013) measured depression using the Center of Epidemiological Studies-Depression scale (Radloff, 1977) and anxiety using the Hospital Anxiety and Depression Scale (Zigmond and Snaith, 1983) in a sample of 479 insomnia patients but reported only stratified summary statistics of the anxiety scores after dichotomization of the depression variable, using a cutoff of 15.5 on the Center of Epidemiological Studies-Depression scale to create two groups with low ($n_0 = 198$) versus mild/high ($n_1 = 281$) amounts of depression.[8] Using Equations [3], [5], and [8], the biserial correlation coefficient can then be computed based on the information provided. The resulting coefficient of $r_b = 0.703$ provides this study's estimate of the association between the underlying continuous variables and can be integrated with the product–moment correlations of studies 1–6. In addition, the coefficient's variance was computed using Soper's exact Equation [12], which yields 0.0012. These values are entered for columns y_i and v_i in row 7 of Table 2.

Finally, Jansson and Linton (2006) report the results of a general population survey among 1936 participants, identifying 212 individuals with insomnia. Depression and anxiety were measured with the Hospital Anxiety and Depression Scale, using a cutoff of 7.5 to dichotomize both variables. The cell frequencies of the resulting 2 × 2 table (with a_i, b_i, c_i, and d_i denoting the respective cell frequencies) are given in Table 2. Based on this information, the tetrachoric correlation coefficient and its (asymptotic) variance can be computed using known methods (e.g., Brown, 1977; Hamdan, 1970; Olsson, 1979; Tallis, 1962), which yields $r_{tet} = 0.576$ with corresponding sampling variance equal to 0.0074. Because the tetrachoric correlation coefficient is an estimate of the product–moment correlation of the underlying continuous variables, it can be integrated with the estimates of studies 1–7, and hence, we enter the estimate and its corresponding sampling variance in row 8 for columns y_i and v_i, respectively.

A forest plot of the findings of the eight studies is shown in Figure 6. Note that the confidence intervals for the individual studies were computed in different ways, depending on the type of measure used. For studies 1–6, the confidence intervals were constructed based on Fisher's *r*-to-*z* transformation (i.e., using the equation $tanh\left[\text{arctanh}\ [r_i] \pm 1.96\sqrt{1/(n_i - 3)}\right]$, where *tanh* is the hyperbolic tangent function and arctanh its inverse). For study 7, which yields a biserial correlation coefficient, we used the variance-stabilizing transformation derived in the present paper (i.e., using Equations [17], [18], and [19]) to construct the confidence interval. Finally, for study 8, we computed a simple Wald-type confidence interval based on the tetrachoric correlation coefficient and its estimated sampling variance (i.e., using the equation $y_i \pm 1.96\sqrt{v_i}$).

The findings of the individual studies can be integrated using standard meta-analytic methods (e.g., Shadish and Haddock, 2009). Applying a random-effects model to the various correlation coefficients (y_i) together with their corresponding sampling variances (v_i) yields an estimated average correlation of $\hat{\mu} = 0.61$ with 95% confidence interval (0.56, 0.67). These results are displayed at the bottom of the forest plot in Figure 6. For comparison purposes, we also added the estimate when only including studies 1–6 in the meta-analysis (i.e., when

[8]To be precise, results were stratified based on three levels of depression (i.e., low, mild, and high), but we collapsed the mild and high groups to obtain an example of a study with a dichotomized variable. Based on the three levels of depression, one could also compute the polyserial correlation coefficient (Bedrick and Breslin, 1996; Olsson *et al.*, 1982), the extension of the biserial correlation to more than two groups, but discussion of this is beyond the scope of the present paper.

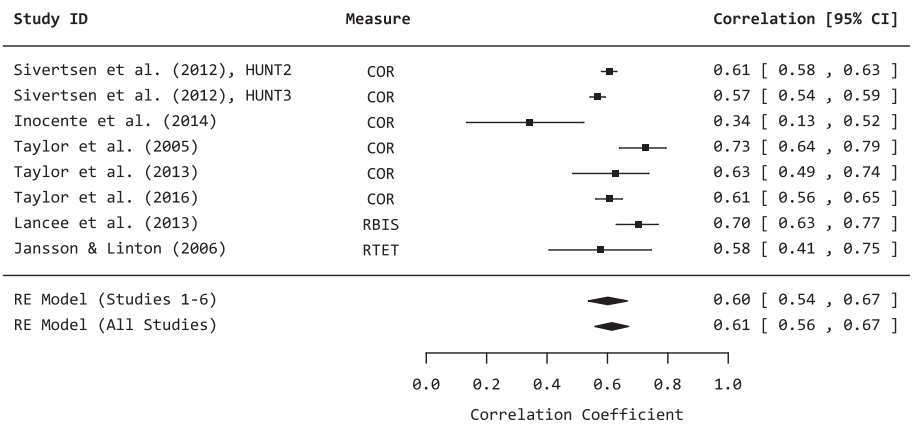

Figure 6. Forest plot based on the data in Table 2.

excluding the two studies not directly reporting a Pearson product–moment correlation coefficient). While the estimate itself is changed very little, some efficiency is lost, leading to a slightly wider confidence interval.

As part of the Supporting Information, we provide the R code to reproduce the results of this illustrative example, using the metafor package for the meta-analysis (Viechtbauer, 2010).

8.6. Final issues

One of the issues discussed earlier is the type of cutoff used for the dichotomization. With only few exceptions, neither the estimates nor their sampling variances were considerably affected by the type of cutoff used in dichotomizing the sample. Thus, although Pearson (1909a) and Soper (1914) derived the biserial correlation coefficient and its sampling variance assuming dichotomization at a hard cutoff, the present study suggests their validity in situations when the cutoff is based on sample percentiles. Conversely, the variance-stabilizing transformation suggested here was derived for the case of an adaptive cutoff point, treating p, q, and $f(z_p)$ as constants. The present results, however, suggest that the transformation can still be applied even when the dichotomization is based on a previously fixed cutoff value rather than a sample percentile.

Instead of using percentiles to define the cutoff for dichotomization, researchers could also decide to dichotomize at $c = \bar{x}$ or $c = \bar{x} \pm zs_x$, that is, at the sample mean or a certain number (z) of standard deviations above or below the sample mean. This approach would represent a hybrid of a hard and an adaptive cutoff, as the cutoff value is then defined *ex-post* based on sample statistics, but does not fix group sizes in finite samples. However, given that the type of cutoff used had only a minor influence on the results, we would not expect qualitatively different results if such an approach were to be used.

One aspect of the simulation study calls for some further discussion. As described earlier, we decided to omit (and replace) iterations where either n_1 or n_0 was less than two, which can happen when a hard cutoff value is used. In practice, dichotomization at a hard cutoff would probably be abandoned as a sensible option when doing so would lead to one of the groups including no observations. It is also unlikely that researchers would report "summary statistics" for a "group" that consists of a single observation. Therefore, cases where there are fewer than two observations in one of the groups are unlikely to be seen in practice in the first place. However, it needs to be kept in mind that the results given are conditional on this case not occurring.

Moreover, it needs to be emphasized that the results given are only applicable when X and Y follow a bivariate normal distribution. In practice, however, one of the marginal distributions may be far from normal (for some empirical evidence that univariate distributions are often non-normal, see Micceri, 1989), this being the very reason why a researcher decides to dichotomize that specific variable in the first place (leaving aside whether this is necessary or the best way to handle the data). Because the marginal distributions of a bivariate normal distribution must also be normal, this would then be an indication that the joint distribution is in fact not bivariate normal. In that case, the biserial correlation coefficient may no longer be (approximately) unbiased, and its sampling variance may be estimated much more inaccurately. The influence of non-normality on the biserial correlation coefficient remains subject to further investigation.

Finally, it is important to emphasize that other statistical artifacts besides artificial dichotomization may be present in a given dataset. For example, measurement error and range restriction in one or both variables of interest will lead to correlation coefficients that are diluted to some extent, requiring other corrections to disattenuate the reported values (for example, Hunter and Schmidt, 2004). Additional work will be needed to develop appropriate methods to estimate the underlying (disattenuated) correlation coefficient of the continuous variables (and its sampling variance) when the results of a study are affected not only by artificial dichotomization but also by additional artifacts.

References

Beck AT, Steer RA, Ball R, Ranieri WF. 1996. Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. *Journal of Personality Assessment* **67**: 588–597.

Becker MP, Clogg CC. 1988. A note on approximating correlations from odds ratios. *Sociological Methods & Research* **16**: 407–424.

Bedrick EJ, Breslin FC. 1996. Estimating the polyserial correlation coefficient. *Psychometrika* **61**: 427–443.

Boas F. 1909. Determination of the coefficient of correlation. *Science* **29**: 823–824.

Bonett DG, Price RM. 2005. Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics* **30**: 213–225.

Borenstein M. 2009. Effect sizes for continuous data. In Cooper H, Hedges LV, Valentine JC (eds.). The Handbook of Research Synthesis and Meta-analysis (pp. 221–235). 2nd edn.New York: Russell Sage Foundation.

Bosco FA, Aguinis H, Singh K, Field JG, Pierce CA. 2015. Correlational effect size benchmarks. *Journal of Applied Psychology* **100**: 431–449.

Brown MB. 1977. Algorithm AS 116: the tetrachoric correlation and its asymptotic standard error. *Journal of the Royal Statistical Society, Series C* **26**: 43–351.

Cohen J. 1983. The cost of dichotomization. *Applied Psychological Measurement* **7**: 249–253.

Cooper H. 2009. Hypotheses and problems in research synthesis. In Cooper H, Hedges LV, Valentine JC (eds.). The Handbook of Research Synthesis and Meta-analysis (pp. 19–35). 2nd edn.New York: Russell Sage Foundation.

Digby PGN. 1983. Approximating the tetrachoric correlation coefficient. *Biometrics* **39**: 753–757.

Fisher RA. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**: 507–521.

Fisher RA. 1921. On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron* **1**: 1–32.

Hamdan MA. 1970. The equivalence of tetrachoric and maximum likelihood estimates of ρ in 2×2 tables. *Biometrika* **57**: 212–215.

Hedges LV, Olkin I. 1985. Statistical Methods for Meta-analysis. San Diego, CA: Academic Press.

Hunter JE, Schmidt FL. 1990. Dichotomization of continuous variables: the implications for meta-analysis. *Journal of Applied Psychology* **75**: 334–349.

Hunter JE, Schmidt FL. 2004. Methods of Meta-analysis: Correcting Error and Bias in Research Findings. 2nd edn. Thousand Oaks, CA: Sage.

Inocente CO, Gustin MP, Lavault S, Guignard-Perret A, Raoux A, Christol N, Gerard D, Dauvilliers Y, Reimão R, Bat-Pitault F, Lin JS, Arnulf I, Lecendreux M, Franco P. 2014. Depressive feelings in children with narcolepsy. *Sleep Medicine* **15**: 309–314.

Jansson M, Linton SJ. 2006. The role of anxiety and depression in the development of insomnia: cross-sectional and prospective analyses. *Psychology and Health* **21**: 383–397.

Koopman RF. 1983. On the standard error of the modified biserial correlation. *Psychometrika* **48**: 639–641.

Lancee J, van den Bout J, van Straten A, Spoormaker VI. 2013. Baseline depression levels do not affect efficacy of cognitive-behavioral self-help treatment for insomnia. *Depression and Anxiety* **30**: 149–156.

Lev J. 1949. The point biserial coefficient of correlation. *Annals of Mathematical Statistics* **20**: 125–126.

Lewis S, Clarke M. 2001. Forest plots: trying to see the wood and the trees. *British Medical Journal* **322**: 1479–1480.

Lipsey MW, Wilson DB. 2001. Practical Meta-analysis. Thousand Oaks, CA: Sage.

Lord FM. 1963. Biserial estimates of correlation. *Psychometrika* **28**: 81–85.

MacCallum RC, Zhang S, Preacher KJ, Rucker DD. 2002. On the practice of dichotomization of quantitative variables. *Psychological Methods* **7**: 19–40.

Maxwell SE, Delaney HD. 1993. Bivariate median splits and spurious statistical significance. *Psychological Bulletin* **113**: 181–190.

Meyer GJ, Finn SE, Eyde LD, Kay GG, Moreland KL, Dies RR, Eisman EJ, Kubiszyn RW, Reed GM. 2001. Psychological testing and psychological assessment: a review of evidence and issues. *American Psychologist* **56**: 128–165.

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

Olsson U. 1979. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**: 443–460.

Olsson U, Drasgow F, Dorans NJ. 1982. The polyserial correlation coefficient. *Psychometrika* **47**: 337–347.

Pearson K. 1900. Mathematical contribution to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, Series A* **195**: 1–47.

Pearson K. 1909a. On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A. *Biometrika* **7**: 96–105.

Pearson K. 1909b. Determination of the coefficient of correlation. *Science* **30**: 23–25.

Pearson K. 1913. On the probable error of a coefficient of correlation as found from a fourfold table. *Biometrika* **9**: 22–27.

Pearson K, Heron D. 1913. On theories of association. *Biometrika* **9**: 159–315.

Pustejovsky JE. 2014. Converting from *d* to *r* to *z* when the design uses extreme groups, dichotomization, or experimental control. *Psychological Methods* **19**: 92–112.
R Core Team. 2014. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
Radloff LS. 1977. The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement* **1**: 385–401.
Raudenbush SW. 2009. Analyzing effect sizes: random-effects models. In Cooper H, Hedges LV, Valentine JC (eds.). The Handbook of Research Synthesis and Meta-analysis (pp. 295–315). 2nd edn.New York: Russell Sage Foundation.
Richard FD, Bond CF, Stokes-Zoota JJ. 2003. One hundred years of social psychology quantitatively described. *Review of General Psychology* **7**: 331–363.
Rosenthal R. 1991. Meta-analytic Procedures for Social Research. Thousand Oaks, CA: Sage.
Schild AH, Voracek M. 2013. Less is less: a systematic review of graph use in meta-analyses. *Research Synthesis Methods* **4**: 209–219.
Shadish WR, Haddock CK. 2009. Combining estimates of effect size. In Cooper H, Hedges LV, Valentine JC (eds.). The Handbook of Research Synthesis and Meta-analysis (pp. 257–277). 2nd edn.New York: Russell Sage Foundation.
Sivertsen B, Salo P, Mykletun A, Hysing M, Pallesen S, Krokstad S, Nordhus IH, Øverland S. 2012. The bidirectional association between depression and insomnia: the HUNT study. *Psychosomatic Medicine* **74**: 758–765.
Soper HE. 1914. On the probable error of the bi-serial expression for the correlation coefficient. *Biometrika* **10**: 384–390.
Tallis GM. 1962. The maximum likelihood estimation of correlation from contingency tables. *Biometrics* **18**: 342–353.
Tate RF. 1954. Correlation between a discrete and a continuous variable: point-biserial correlation. *Annals of Mathematical Statistics* **25**: 603–607.
Tate RF. 1955a. Applications of correlation models for biserial data. *Journal of the American Statistical Association* **50**: 1078–1095.
Tate RF. 1955b. The theory of correlation between two continuous variables when one is dichotomized. *Biometrika* **42**: 205–216.
Taylor DJ, Bramoweth AD, Grieser EA, Tatum JI, Roane BM. 2013. Epidemiology of insomnia in college students: relationship with mental health, quality of life, and substance use difficulties. *Behavior Therapy* **44**: 339–348.
Taylor DJ, Lichstein KL, Durrence H, Reidel B, Bush AJ. 2005. Epidemiology of insomnia, depression, and anxiety. *Sleep* **28**: 1457–1764.
Taylor DJ, Pruiksma KE, Hale WJ, Kelly K, Maurer D, Peterson AL, Mintz J, Litz BT, Williamson DE, STRONG STAR Consortium. 2016. Prevalence, Correlates, and Predictors of Insomnia in the US Army Prior to Deployment. *Sleep*. in press.
Venables WN, Ripley BD. 2002. Modern Applied Statistics with S. 4th edn.New York: Springer.
Viechtbauer W. 2010. Conducting meta-analyses in R with the meta for package. *Journal of Statistical Software* **36**: 1–48.
Wicherts JM, Borsboom D, Kats J, Molenaar D. 2006. The poor availability of psychological research data for reanalysis. *American Psychologist* **61**: 726–728.
Yule GU. 1912. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society* **75**: 579–652.
Zigmond AS, Snaith RP. 1983. The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica* **67**: 361–370.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.