Meta-analyzing dependent correlations with correction for artifacts that multiplicatively attenuate the true correlation

Shu Fai Cheung¹ · Darius K.-S. Chan² · Rong Wei Sun¹

Published online: 22 August 2018
© Psychonomic Society, Inc. 2018

Abstract

Previous procedures for meta-analyzing dependent correlations have been found to overestimate or underestimate the true variation in effect sizes. Samplewise-adjusted procedures have been shown to perform better than simple within-study means when meta-analyzing dependent correlations. However, such procedures cannot be applied when correction for artifacts such as unreliability is desired. In the present study, we extended the procedures to correct for attenuation due to artifacts when meta-analyzing dependent correlations. Monte Carlo simulation was conducted in order to examine conditions with various degrees of dependence, degrees of heterogeneity, sample sizes, and numbers of studies, among other factors. The previous procedures, including the samplewise-adjusted procedures without correction, yielded biased point estimates and confidence intervals with low coverage probabilities of the population mean correlation and degree of heterogeneity. More importantly, the bias and undercoverage of the confidence interval increased with the mean sample size and number of studies in many conditions. The new samplewise-adjusted procedures with correction for attenuation yielded negligible biases when estimating the mean population correlation, even in the presence of dependent correlations. Given that the need for correction for attenuation due to artifacts is becoming more recognized in meta-analysis, our findings highlight the importance of such considerations when meta-analyzing dependent correlations. Conditions under which these procedures can be further improved are also discussed.

Keywords Meta-analysis · Dependent effect sizes · Correction for artifacts

Meta-analysis has become a major approach for summarizing previous studies in applied psychology (e.g., Colquitt et al., 2013, on justice; de Wit, Greer, & Jehn, 2012, on intragroup conflict; Wang, Waldman, & Zhang, 2014, on team effectiveness). In a meta-analysis, the effect sizes from comparable studies are quantitatively combined in order to yield estimates of the mean effect, the “true” variability of the effect not due to sampling error and other artifacts, and the effects of study characteristics (usually called *moderators* in meta-analysis) on the magnitude and sign of the effect. Various schools of techniques have been proposed for use in conducting meta-analysis. They differ in the assumptions made, the degree and

nature of the true variation of effect sizes, and the practical constraints that are common in meta-analyses (e.g., Hedges & Vevea, 1998; Hunter & Schmidt, 2004).

One of the common problems in meta-analysis is dependence among the effect sizes. It is common that more than one effect size was reported in a study. The effect sizes from the same study can be dependent for various reasons (described in the next section). This is a problem because most common procedures for meta-analysis assume that all effect sizes being summarized are independent (e.g., Hedges & Olkin, 1985; Hedges & Vevea, 1998; Hunter & Schmidt, 2004). If the dependence among some subsets of effect sizes is ignored, the standard error of the weighted average effect size and the estimated true variation across studies, also called the *degree of heterogeneity*, will be biased (Cheung & Chan, 2004, 2008, 2014; Martinussen & Bjørnstad, 1999; Viswesvaran, Sanchez, & Fisher, 1999). The consequences can be substantial, because the estimated true variation influences the extent to which we believe that the variation in effect sizes needs to be explained by moderators. Failure to detect heterogeneity results in failure to find moderators that influence an effect, such as the culture in

✉ Shu Fai Cheung
sfcheung@umac.mo

¹ Department of Psychology, Faculty of Social Sciences, University of Macau, Avenida da Universidade, Taipa, Macau, China

² Department of Psychology, Chinese University of Hong Kong, Hong Kong, China

which a training strategy is effective or conditions in which a personality trait has a strong effect on job performance.

In the present article, we will briefly review selected procedures proposed to handle dependent effect sizes. We will then focus on the samplewise-adjusted procedure (Cheung & Chan, 2004) and extend this procedure to correct for artifacts in meta-analyzing dependent effect sizes. We will also present a simulation study to compare the performance of the extended samplewise-adjusted procedure with a few other procedures.

The scope of the present article

There are two common ways that effect sizes are grouped or clustered. First, several studies, each contributing only one effect size, were conducted by the same research team and/or published in the same article. Second, several effect sizes were reported for the same study, all from the same sample. For example, a study may report several correlations between job performance and emotional intelligence. There are two common situations in which these correlations are dependent. The first one is when all the correlations are based on the same measures of job performance and emotional intelligence but are measured at different time points. To the extent that scores from the same measure at different times are correlated, the correlations are dependent (analogous to multiple-endpoint studies, Gleser & Olkin, 2009, or multiple-time-point studies, Borenstein, Hedges, Higgins, & Rothstein, 2009). Another common situation is when the correlations are computed between different measures of one construct (e.g., emotional intelligence) and the same measure of another construct (e.g., job performance). These correlations are dependent because they share a common variable (analogous to multiple-treatment studies, Gleser & Olkin, 2009, or multiple-outcome studies, Borenstein et al., 2009).

Strictly speaking two variables can be dependent but uncorrelated. However, in meta-analyses, the most common form of dependence is being correlated, and so we focus on correlated effect sizes in the present article. To avoid confusing terms such as *correlated correlations*, we use the term *dependent effect sizes* to refer to correlated effect sizes.

We believe that the degree of dependence in multiple-outcome or -time-point studies is higher than that due to the effect sizes due to findings from the same study or research team. Moreover, it is a common practice to treat two or more effect sizes from the same article (i.e., produced by the same researcher or research teams) as independent if they come from independent studies. Therefore, we will assume that the unit of analysis in a meta-analysis is the *sample*, and that dependent effect sizes are grouped by samples. Because it is common in meta-analyses to use the term *study* to refer to the unit of analysis, we will use the term *study* instead of *sample* throughout the article.

Moreover, our illustration and simulation study will focus on meta-analyzing sample correlations. Therefore, we will also restrict our discussion to effect sizes coming from the same study and having identical or nearly identical sample sizes (e.g., minor sample size differences that, if they exist, are due to missing data). We will not investigate complicated scenarios in which several effect sizes, with very different sample sizes, are dependent because they shared the same control group, such as an experiment in which several treatment groups are compared to a control group (see Gleser & Olkin, 2009, on how to handle these kinds of dependent effect sizes).

A brief review of major approaches

It is beyond the scope of the present article to give a comprehensive review of major approaches in handling dependent effect sizes (see M. W. L. Cheung, 2015, chapter 6, for an in-depth review). We will only review the essential characteristics of three major approaches: the analytic approach, the three-level modeling approach, and the sample-wise approach. We will then focus on the sample-wise approach, which we will extend to correct for artifacts.

In the *analytic* approach, a regression model is used to model the effect sizes, as in the meta-analysis of independent effect sizes, with the number of units being equal to the number of effect sizes, even if some of the effect sizes came from the same study. There is only one random effect: the true variation between the effect sizes. For example, the true variation of ten attitude–behavior correlations from ten independent studies is assumed to be the same as the true variation of ten attitude–behavior correlations from the same study. The dependence within groups of effect sizes is modeled by specifying nonzero covariances for the residuals. For example, Becker's (1992) generalized least squared meta-analysis and the procedures by Gleser and Olkin (2009) for experimental studies include the covariances among dependent effect sizes in the model. We only need to specify how the effect sizes, for whatever reasons, are correlated. This approach is flexible and can handle forms of dependence that can be treated as covariances among residuals. However, these covariances are assumed to be known, which is rarely the case. The robust variance estimation (RVE) method was proposed partly to address this limitation (Hedges, Tipton, & Johnson, 2010; Tipton, 2013, 2015). Similar to the generalized least squares approach, the inter-effect-size dependence is acknowledged, and a working value for the unknown covariances is used in the procedure. However, the results of the RVE method are supposed to be insensitive to the choice of the imputed value. This approach is promising but has yet to be examined in more situations.

The second approach, the *three-level modeling* approach, extends the two-level meta-analytic model (Raudenbush & Bryk, 2002) by including the grouping of dependent effect sizes as an additional level. When each study contributes only one effect size and so each group has only one unit, the model can be reduced to a two-level model. This approach can model the dependence among effect sizes due to clustering (Cheung, 2013, 2015; Marsh, Bornmann, Mutz, Daniel, & O'Mara, 2009; van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013). Unlike in the analytic approach, there are two random effects, those between studies and within studies. For example, the true variation of mean attitude–behavior correlations between studies is the between-studies random effect, and the true variation of attitude–behavior correlations within each study is the within-study random effect. Also, different from the analytic approach, the residuals are assumed to be independent. The dependence within each group of effect size is equal to the between-studies random effect (M. W. L. Cheung, 2015).

The third approach, the *sample-wise* approach, is a two-step approach. First, each group of dependent effect sizes is replaced by the mean of them, denoted as the within-study mean (this mean can be a weighted mean, as in we use in the Appendix for computing the mean with correction for artifacts, or an unweighted mean, as in the computation of the synthetic mean of effect sizes discussed by Borenstein et al., 2009). These means are then used in place of the groups of effect sizes for a subsequent meta-analysis using meta-analytic procedures that assume independent effect sizes. This method is justifiable when the effect sizes come from the same sample and the goal is to form a summary effect from these dependent effect sizes (Borenstein et al., 2009). For example, several attitude–behavior correlations, for different behaviors, are reported for the same sample, and we want to compute the summary association between attitude and behavior for this sample. Because the within-study mean provides more information than does one effect size from a study of the same sample size, the samplewise-adjusted procedure has been proposed to compute the effective sample size of the within-study mean on the basis of the estimated degree of dependence among effect sizes (Cheung & Chan, 2004, 2008). Similar to the analytic approach, covariances among effect sizes are allowed to be different (Cheung & Chan, 2008). For example, for a study with ten attitude–behavior correlations, some pairs of correlations may have higher covariances than other pairs. Similar to the three-level modeling approach, the true variation of the means of effect sizes is estimated. The true variation estimated in the second step in the sample-wise approach corresponds to the between-study random effect in the three-level modeling approach. Unlike in the analytic approach and the three-level modeling approach, all dependent effect sizes from the same study are assumed to be homogeneous, or at least have the same direction in the

population and the within-study heterogeneity is negligible. For example, all population attitude–behavior correlations within a study are assumed to have the same value.

None of these three approaches is unconditionally better than the others. As is apparent from the brief review, the three approaches make different assumptions about the way the effect sizes are dependent and the way they vary within and between studies. None of them is a special case of one of the others. The analytic approach and the sample-wise approach allow for variation in dependence (within study covariances), whereas the three-level modeling approach assumes that the degree of dependence is the same across studies. The analytic approach and three-level modeling approach allow for true variation of the effect sizes within groups of effect sizes, whereas the sample-wise approach assumes that all effect sizes within a group is homogeneous. The analytic approach and the three-level modeling approach limit researchers as to the way to do a meta-analysis, because they are meta-analytic procedures that handle dependent effect sizes, whereas the sample-wise approach is a two-step approach that reduces groups of effect sizes into means with adjusted sampling variances, with no limitation on the meta-analytic procedure to be used on these means. In sum, all three approaches have their pros and cons, and selecting the procedure depends on the assumptions that are tenable and the meta-analytic approach to be adopted, among other factors.

In the present article, our focus is on correction for artifacts in meta-analysis. To the best of our knowledge, no established procedure is available to correct for unreliability when using the analytic approach (including the RVE method) and the three-level modeling approach. Moreover, as we reviewed above, these three approaches represent different classes of approaches intended for different situations. For example, when the effect sizes within each study show negligible heterogeneity (i.e., the within-study true variation is not equal to between-study true variation, violating an assumption of the analytic approach) and the within-study effect sizes are correlated (violating the assumption of uncorrelated residuals in the three-level modeling approach; Cheung & Chan, 2014), the sample-wise approach is more appropriate than the other two approaches. Therefore, we consider only the sample-wise approach in the present study, in which we propose a way to correct for artifacts. We will show that samplewise-adjusted procedures (Cheung & Chan, 2004, 2008) can be extended to incorporate correction for artifacts. Therefore, in the rest of this article we will focus on the sample-wise approach.

Correcting for artifacts when meta-analyzing correlations

In meta-analysis, the goal is to find the true relation between variables—that is, the correlation at the construct level

(Schmidt & Hunter, 2015). However, aside from sampling error, the observed correlations can be distorted by imperfections in a study, which Schmidt and Hunter (1977, 2015) called *artifacts* in order to stress that they are artifactual errors. In addition to sampling error, Schmidt and Hunter (2015) identified ten other artifacts usually found in correlational studies, such as unreliability (error in measurement), dichotomization, range restriction, and imperfect validity. Most of the time (though not always; see, e.g., Iacobucci, Posavac, Kardes, Schneider, & Popovich, 2015, on median split) the presence of these artifacts tends to reduce (attenuate) the observed correlation, resulting in underestimation of the true relation between the variables of concern.

Correcting for artifacts is important in meta-analysis for two reasons. First, such a correction leads to less biased estimates of the mean correlation (Schmidt & Hunter, 2015; Schmidt, Le, & Oh, 2009). Second, studies vary in their artifacts (Schmidt & Oh, 2016; see also Peterson, 1994, for a meta-analysis of Cronbach's alphas). For example, reliability is sample-dependent, and thus varies across studies. Ignoring variation in artifacts such as unreliability results in biased estimation of the degree of heterogeneity of population correlations. In the following section, we will first briefly present the details of the samplewise-adjusted procedures, and then propose a modification that allows for correction for artifacts.

We will focus on artifacts that attenuate a correlation by multiplying it, such as unreliability and dichotomization. For example, if the reliability of a variable is .81 and the true correlation is .60, the attenuated correlation is $(\sqrt{.81}) (.60) = .54$. If the reliability of one variable is .81, the construct validity of the other variable is .70, both variables were dichotomized at the medians (the resulting attenuation is .80; Schmidt & Hunter, 2015), and the true correlation is .80, the attenuated correlation is $(\sqrt{.81}) (.70) (.80) = (.504) (.80) = .4032$. The total attenuation of these three sources of artifacts (unreliability, imperfect construct validity, and median split) is .504 (see Schmidt & Hunter, 2015, chapter 3, for other examples). Therefore, it should be noted that although we choose to use unreliability as an example of an artifact in our numerical illustration and simulation study, our discussion and all of the findings derived can be generalized to other artifacts and to combinations of such artifacts, as long as they attenuate the true correlation in the aforementioned multiplicative manner.

Samplewise-adjusted procedures

The two samplewise-adjusted procedures proposed by Cheung and Chan (2004, 2008) are the adjusted individual and adjusted weighted procedures. The samplewise-adjusted individual procedure involves two steps: estimating the degree of dependence of the effect sizes from a study and computing

an effective sample size that reflects the weight of the within-study mean of these dependent effect sizes. Assume that we have k studies and that the i th study had p_i effect sizes ($p_i \geq 1$). The j th and l th effect sizes in the i th study are denoted by r_{ij} and r_{il} , respectively. The correlation between the j th and l th effect sizes (i.e., between r_{ij} and r_{il}) is denoted by b_{ijl} . In the first step, the population mean inter-effect-size correlation (b_i) for the i th sample with multiple effect sizes ($p_i \geq 2$) is estimated by the following equation:

$$\hat{b}_i = 1 - \frac{S_{ES_i}^2}{\sigma_{ei}^2}$$

where σ_{ei}^2 is the average theoretical sampling variance for the i th study and $S_{ES_i}^2$ is the observed variance of the dependent effect sizes (Cheung & Chan, 2008). If the effect sizes from a study are independent (all $b_{ijl} = 0$), the observed variance and the average theoretical sampling variance are expected to be equal, and therefore $\hat{b}_i$ is expected to be zero. If the effect sizes from a study are correlated (i.e., one or more b_{ijl} in the i th study are nonzero), the within-study variance is expected to be smaller than the average theoretical sampling variance, and consequently it is expected that $\hat{b}_i$ will be greater than zero.

In the second step, the sampling variance of the within-study mean is estimated by $C_i \sigma_{ei}^2$, where the correction factor C_i is given by

$$C_i = \frac{1 + (p_i - 1)\hat{b}_i}{p_i}$$

If the estimate of sampling variance is a function of the sample size and sample effect size (y), such as $f(y)/(n_i + d)$, where d is a constant and $f(y)$ is a function of the sample effect size y , then C_i can be used to compute the effective sample size $n'_i = (n_i + d) / C_i - d$. The effective sample size can then be used as the sample size of the within-study mean for further analysis. If sample size weighting is used as in the Hunter–Schmidt approach, then the effective sample size for the within-study mean is n_i / C_i . Therefore, within-study means can be used in the main meta-analysis, whether sampling variances or sample sizes are used in the procedures or software packages are used for weighting.

The samplewise-adjusted weighted procedure modifies the samplewise-adjusted individual procedure by computing the weighted average of $\hat{b}_i$ for all studies with dependent effect sizes and using this weighted average ($\bar{b}$) to adjust the sampling variances and sample sizes of all within-study means. This is based on the logic of the Hunter–Schmidt approach. The weighted average correlation, being more stable (i.e., smaller sampling variance) than the individual sample correlations, can provide a better estimate of the sampling variance of each sample correlation. This rationale can justify the use of

sample size weighting in the Hunter–Schmidt approach. Similarly, the weighted average ($\bar{b}$) can serve as a better basis for the correction than the individual $\hat{b}_i$, especially when each group of dependent effect sizes only has a few effect sizes. Cheung and Chan (2008) demonstrated empirically that the samplewise-adjusted weighted procedure yields less biased estimates of the degree of heterogeneity than does the samplewise-adjusted individual procedure, even when the degree of dependence (b_i) varies across studies.

Previous simulation studies have demonstrated that the samplewise-adjusted procedures provide less biased estimates of the degree of heterogeneity, especially when the degree of dependence is high (Cheung & Chan, 2004, 2008). In a comparison with three-level modeling, Cheung and Chan (2014) also found that the samplewise-adjusted procedures were less biased than three-level modeling when the degree of dependence was high and the assumption of negligible within-study heterogeneity was met.

Meta-analyzing dependent correlations with correction for artifacts

In this section, we present a modification of the samplewise-adjusted procedures to accommodate correction for artifacts. We focus our discussion on meta-analyzing correlations because correction for artifacts is usually applied to this kind of effect size measures. The derivation of the formulas is presented in the Appendix. It is shown that the samplewise-adjusted procedure can be extended to disattenuated correlations. First, the disattenuated correlation is computed by

$$r_{Cij} = r_{ij}/A_{ij},$$

where A_{ij} is the attenuation factor for the j th correlation of the i th study. For example, if a correlation is to be corrected for unreliability in one of the variables, A_{ij} is the square root of the reliability (Hunter & Schmidt, 2004; Schmidt & Hunter, 2015; Spearman, 1904). If a correlation is to be corrected for dichotomization at the median in another variable, A_{ij} is .80 (Schmidt & Hunter, 2015). If a correlation is to be corrected for both unreliability of one of the variables and a median split in another variable, $A_{ij} = (\text{square root of the reliability})(.80)$.

Second, the within-study weighted mean of the disattenuated correlations for the i th study is computed by Eq. A4:

$$r_{Ci} = \sum_{j=1}^{p_i} A_{ij}^2 r_{Cij} / \left(\sum_{j=1}^{p_i} A_{ij}^2 \right)$$

This equation assumes that the sample size for all correlations in the i th study is the same, which is usually the case for dependent correlations coming from the same sample. If the sample sizes are slightly different for the correlations from the i th study, usually due to missing data, then the mean,

minimum, or maximum sample size can be used, depending on whether we want to use all information, be conservative, or be liberal. In practice, the sample sizes usually are similar, if not identical, so using whichever working value of sample size should have minimal impact on the meta-analytic results.

In the meta-analysis of independent correlations with correction for artifacts, the attenuation factor of a correlation is used to compute the sampling variance (σ_{ei}^2/A_i^2). For a within-study weighted mean of p_i disattenuated correlations, there are p_i attenuation factors. An effective attenuation factor can be computed such that the sampling variance of a within-study weighted mean disattenuated correlation is still of a similar form. We will present two ways to compute this effective attenuation factor.

For the adjusted individual procedure, as shown in the Appendix, the effective attenuation factor for the within-study weighted mean of disattenuated (r_{Ci}) can be computed by Eq. A6,

$$A_i = \left(\sum_{j=1}^{p_i} A_{ij}^2 \right) / \sqrt{\sum_{j=1}^{p_i} A_{ij}^2 + \sum_{j=1}^{p_i} \sum_{l=1, l \neq j}^{p_i} A_{ij} A_{il} \hat{b}_i}$$

where $\hat{b}_i$ is the same estimate in the adjusted individual procedure, without correction for attenuation.

For the adjusted weighted procedure, the weighted mean of $\hat{b}_i, \hat{b}$, is used for all studies with multiple effect sizes (Eq. A7),

$$A_i = \left(\sum_{j=1}^{p_i} A_{ij}^2 \right) / \sqrt{\sum_{j=1}^{p_i} A_{ij}^2 + \sum_{j=1}^{p_i} \sum_{l=1, l \neq j}^{p_i} A_{ij} A_{il} \hat{b}}$$

The only difference between the adjusted weighted procedure and the adjusted individual procedure is how the sampling variance of a within-study weighted mean of disattenuated correlation is computed.

The two procedures basically involve three steps: (1) for the i th study with dependent effect sizes, replace the p_i dependent correlations by the within-study weighted mean of corrected correlation, r_{Ci} ; (2) use the weighted average uncorrected correlation to compute the sampling variance, $\sigma_{ei}^2 = (1 - \bar{r}_i^2)^2 / (n_i - 1)$; and (3) weigh the within-study weighted mean of corrected disattenuated correlation (r_{Ci}) by σ_{ei}^2/A_i^2 . The two procedures differ on how A_i is computed. After all groups of dependent disattenuated correlations are reduced to within-study weighted mean disattenuated correlations, they can be meta-analyzed together with other independent disattenuated correlations, treating each within-study mean disattenuated correlations as a usual disattenuated correlations with sample size n_i and attenuation factor A_i .

If a program for meta-analyzing correlations without correction for artifacts is used, an effective sample size can be computed for each disattenuated correlation as the input for

the program, whether it is a within-study weighted mean or the only correlation from a study. The effective sample size is given by $n'_i = (n_i - 1)A_i^2 + 1$, where A_i is either the effective attenuation factor (A'_i) of a within-study mean or the usual attenuation factor of a disattenuated correlation (A_i).

These two procedures are more general than the sample-wise procedures proposed by Cheung and Chan (2004, 2008). If the correlations are not corrected for attenuation, all A_{ij} will be equal to 1, and $1/A_i^2$ will be equal to the correction factor C_i in Cheung and Chan's procedure. In other words, Cheung and Chan's procedures can be considered as special cases of the modified procedures proposed here.

Two alternative procedures, sample-wise- n and sample-wise- np , should be mentioned. As we discussed above, the two samplewise-adjusted procedures essentially compute an effective sample size that takes into account the dependence among effect sizes. The sample-wise- n procedure and the sample-wise- np procedure represent two special cases. The sample-wise- n procedure uses the original sample size to weigh the within-study weighted mean, regardless of the number of effect sizes. This choice of sample size ignores the extra information provided by different effect sizes, despite their dependence. The sample-wise- np procedure, on the other hand, uses the original sample size (n) multiplied by the number of effect sizes (p) to weigh the within-study disattenuated mean. This choice of sample size treats all the within-study effect sizes as independent, thus ignoring the degree of dependence. Previous studies have shown that the sample-wise- n procedure tended to underestimate the degree of heterogeneity, whereas the sample-wise- np procedure tended to overestimate the degree of heterogeneity (Cheung & Chan, 2004, 2008, 2014). Following the practice in previous studies of sample-wise procedures, these two procedures were also included for comparison.

A numerical example

To illustrate how the modified sample-wise procedures can be used, we reanalyzed the numerical example published in Cheung and Chan (2014). The dataset consists of 20 correlations from 12 studies (half of them having two or three correlations). For illustration, we randomly generated an attenuation factor for each correlation. The modified dataset can be obtained from https://osf.io/2q3w7/?view_only=3b8a2d7605764352b7e6608bb467507b. The R functions reported in Cheung and Chan (2014) have been modified and included in an R package (*samplewiseart*). The R package and the scripts used in the following example can be obtained from the link above. The detailed instructions to use the function can be found in Cheung and Chan (2014) and also the documentation in the package. In the following discussion, we will only highlight results relevant to correction for artifacts.

The function *MADependentESArtifact* accepts a dataset with five variables: study ID (unique for each study), correlation ID (unique for each correlation in a study), sample size (which should be similar or even identical for correlations within a study), uncorrected correlation, and attenuation factor. The function needs only one column for the attenuation factor. This is because artifacts to be corrected for are converted to attenuation factors and then multiplied to form one attenuation factor. For example, if the reliability coefficient of one variable is .80 and the construct validity coefficient of another variable is .90, and if both variables are dichotomized at their medians, the attenuation factor is $(\sqrt{.80}) (.90) (.80) (.80)$, or about .52.

The function generates a dataset with the number of rows equals to the number of studies. For a study with more than one correlation, the within-study disattenuated mean correlation corrected for artifacts and its sampling variance for each of the samplewise-adjusted procedures are computed. These columns can then be used by other meta-analytic software packages that accept effect sizes and user-calculated sampling variances as the input. The resulting study-level dataset using the six procedures examined in our subsequent simulation study is presented in Table 1. In general, the sampling variance of a corrected correlation is larger than the sampling variance before correction.

We selected the *metafor* package to illustrate how to use the output in Table 1 in other meta-analysis software packages. We used the *rma* function and selected the Hedges–Vevea method. The results are presented in Table 2. To compare the impact of correction for artifacts, we also repeated the analysis without this correction (if the attenuation factor is not specified to *MADependentESArtifact*, it will assume that the attenuation factor is 1 for all correlations and will implement the samplewise-adjusted procedures without correction presented in Cheung & Chan, 2014). As expected, the weighted mean correlations with correction were higher than those without correction (.74 vs. .61 to .62). The estimated random effect (in *SD*) was reduced to zero if corrected for artifacts, and was .0394 to .0496 without correction. Even though the *Q* test was nonsignificant for all procedures except for sample-wise- np , this should be expected because of the small number of studies, which would lead to low statistical power.

As a final remark on this example, the degree of heterogeneity in this example was smaller after correction. This suggests that the variation in the uncorrected correlations was mainly due to the variation in artifacts. However, in practice the degree of heterogeneity can also increase after correction, because the procedures take into account *both* attenuation *and* the degree of dependence. On the one hand, previous studies (e.g., Cheung & Chan, 2004, 2008) have shown that ignoring the degree of dependence can result in either underestimation or overestimation of the degree of heterogeneity, depending on the choice of the sample size (underestimation for the sample-wise- n procedure, and overestimation for the sample-

Table 1 Study-level dataset using the samplewise-adjusted procedures with correction for artifacts

sid	es.mean	n	p	art.mean	es.dis.mean	es.dis.mean.svar					
						Sn	SAdjInd	SAdjInd_C	SAdjWt_C	SAdjWt	Sn _p
1	.69	97	1	.96	.72	.0041	.0040	.0044	.0044	.0040	.0039
2	.57	210	1	.77	.74	.0019	.0018	.0031	.0031	.0018	.0018
3	.58	176	2	.81	.71	.0022	.0018	.0027	.0024	.0015	.0011
4	.62	144	2	.83	.75	.0027	.0016	.0023	.0028	.0019	.0013
5	.63	131	2	.83	.76	.0030	.0021	.0031	.0030	.0021	.0014
6	.70	108	3	.87	.81	.0037	.0034	.0045	.0029	.0022	.0012
7	.44	107	1	.53	.83	.0037	.0036	.0129	.0128	.0036	.0035
8	.65	217	3	.85	.77	.0018	.0013	.0018	.0015	.0011	.0006
9	.62	97	1	.74	.83	.0041	.0040	.0073	.0072	.0040	.0039
10	.62	94	1	.82	.76	.0042	.0042	.0062	.0061	.0041	.0040
11	.58	136	1	.90	.65	.0029	.0029	.0035	.0035	.0028	.0028
12	.65	121	2	.96	.67	.0033	.0016	.0017	.0025	.0023	.0015

sid: study ID; es.mean: within-study mean correlation; *n*: sample size; *p*: number of correlations from a study; *np*: sample size multiplied by the number of correlations; art.mean: the unweighted mean of attenuation factor (for reference only; for $p > 1$, es.mean/art.mean can be different from es.dis.mean); es.dis.mean: the within-study mean of correlations corrected for artifacts; es.dis.mean.svar: the estimated sampling variance of the corrected correlations, adjusted by one of the six procedures; Sn: sample-wise-*n*, SAdjInd: samplewise-adjusted individual; SAdjInd_C: samplewise-adjusted individual with correction for artifacts; SAdjWt_C: samplewise-adjusted weighted with correction for artifacts; SAdjWt: samplewise-adjusted weighted; Sn_p: sample-wise-*np*.

wise-*np* procedure). On the other hand, ignoring the variation in attenuation factor results in overestimation of the degree of heterogeneity (Schmidt & Hunter, 2015). Therefore, whether the estimated degree of heterogeneity will decrease or increase after taking into account these two factors depends on the joint impact of these factors in a particular meta-analysis.

Simulation study

To empirically investigate the performance of the two samplewise-adjusted procedures with correction for artifacts,

we conducted a Monte Carlo simulation study, varying the degree of heterogeneity and degree of dependence, in addition to other common factors such as the number of studies and mean sample size. We investigated how accurate the procedures are in estimating the true population mean correlation when both correction for artifacts and dependence are taken into consideration. We expected that even the samplewise-adjusted procedures would result in biased estimates of the population mean correlation when attenuation was ignored. We also expected that the new samplewise-adjusted procedures with correction for artifacts would outperform previous procedures when some correlations were dependent and the correlations were attenuated by artifacts.

Table 2 Meta-analysis results using the adjusted sampling variances, with the Hedges–Vevea method implemented by the *metafor* package

Method	Estimated mean correlation	Standard error	p	95% Confidence interval	τ^2	τ	<i>Q</i> Statistic	<i>Q p</i>
With Correction								
Samplewise-Adjusted Individual	.74	.0165	.00	.70–.77	.0000	.0000	8.44	.674
Samplewise-Adjusted Weighted	.74	.0163	.00	.71–.77	.0000	.0000	8.48	.669
Without Correction								
Sample-wise- <i>n</i>	.61	.0195	.00	.57–.65	.0016	.0394	15.07	.179
Samplewise-Adjusted Individual	.61	.0193	.00	.58–.65	.0020	.0447	16.38	.128
Samplewise-Adjusted Weighted	.62	.0193	.00	.58–.65	.0021	.0455	17.80	.086
Sample-wise- <i>np</i>	.62	.0191	.00	.58–.66	.0025	.0496	22.37	.022

p: *p* value of testing the estimated mean correlation against zero; τ^2 : estimated random effect variance of correlations; τ : estimated random effect standard deviation of correlations (the square root of τ^2); *Q* statistic: the test of heterogeneity; *Q p*: *p* value of the *Q* statistic.

Previous literature had suggested that the major problem due to dependent effect sizes is estimating the degree of heterogeneity. Therefore, we also examined this aspect of performance. Specifically, we expected that, due to ignoring attenuation, the previous samplewise-adjusted procedures would result in biased estimates of the degree of heterogeneity when some correlations are dependent and the correlations are attenuated by artifacts. We also expected that the new samplewise-procedures with correction for artifacts would outperform previous procedures and result in less biased estimates of the degree of heterogeneity.

In this simulation study, six sample-wise procedures were compared in their performance. In addition to the two samplewise-adjusted procedures with correction for attenuation, we examined four existing sample-wise procedures, all without correction for attenuation: the two samplewise-adjusted procedures proposed by Cheung and Chan (2004), and sample-wise- n and sample-wise- np . The last two procedures represent the two extremes in the choice of sample size for the within-study weighted mean and can serve as benchmarks for the samplewise-adjusted procedures. Because correlation is a common measure of effect sizes in applied research and previous simulation studies of the sample-wise procedures, correlation is used in the simulation study.

Method

Data generation and analysis

We adopted the procedures used in Cheung and Chan (2014). R version 3.5.0 was used to conduct the simulation study (R Core Team, 2018). For each replication, after specifying the population parameters (described below), random raw scores were generated for each study, and then the sample Pearson's r was computed as the effect size for this study. If a study contributed more than one effect size, more than one set of raw scores were generated (dependence controlled as described below). Each condition had 2,000 replications. The R scripts, adapted from Cheung and Chan (2014), were used to implement the sample-wise procedures. The metafor package (Viechtbauer, 2010, version 2.0-0) was used to conduct the meta-analysis using the Hedges approach (method="HE" in the rma function), because the sample-wise procedures can compute the sampling variance of each within-study weighted mean of corrected correlations directly. The confidence interval of metafor for the degree of heterogeneity by Q-profile method was used (Viechtbauer, 2007). Four aspects of meta-analysis were examined: (a) bias in point estimates of the mean population unattenuated correlations (ρ), (b) the empirical coverage probabilities of the 95% confidence intervals of the mean population unattenuated correlations, (c) bias in point estimates of the degree of heterogeneity measured by the standard deviation of the population unattenuated

correlations (τ), and (d) the empirical coverage probabilities of the 95% confidence intervals of the estimated degree of heterogeneity.

Factors manipulated

Mean population correlation Three mean population correlations, .10, .30, and .50, were examined. They corresponded to small, medium, and large effect sizes according to Cohen's (1988) rule of thumb for correlation coefficients.

Degree of heterogeneity (τ) We examined three degrees of heterogeneity: standard deviation of the population correlations equal to 0 (the fixed effect condition with population correlation fixed across studies), .05, and .10. On the basis of Cohen's rule of thumb for small, medium, and large correlations, the difference between categories is .20. The two values of standard deviation were chosen to cover two cases: one-fourth (.05) of this difference and half of this difference (.10). If the mean population correlation is .30 (medium) and the standard deviation is .05, correlations within two SD s of the mean population correlation can still be considered as medium correlations. This case represents the situation in which most correlations fall at the same level (small, medium, or large) as the mean correlation. If the mean population correlation is .30 and the SD is .10, correlations more than one SD below or above the mean population correlation are closer to the neighboring levels (small for .10 and large for .50) than to the medium correlation. This case represents the situation in which the population correlations span two or more levels.

In the random-effects conditions with τ not equal to 0, a beta distribution was used as in Cheung and Chan (2008), such that the population correlations are randomly distributed with known mean and standard deviation within an interval. To prevent the occurrence of extremely high correlation after correction for unreliability, we imposed an upper bound of .80, which is three standard deviations above .50, the largest mean correlation examined, if the standard deviation is .10. The lower bound of the distribution of population correlations was similarly set to $-.80$, such that the distribution is not extremely skewed when the mean correlation is .10 and the standard deviation is .10. The conditions covered are more general than those in previous studies, in which the population correlations were restricted to have the same sign (e.g., nonnegative).

Mean attenuation According to the quantitative review of Cronbach's alpha by Peterson (1994), the mean of Cronbach's alpha in behavioral research was approximately .77 in the studies they reviewed. Therefore, we chose .70 and .80 as the mean reliability coefficients in the present study. The corresponding attenuation factors when the means are .70 and .80, which are the square roots of the reliability

coefficients, are approximately .837 and .894. These are the two levels of mean attenuation we used in the simulation.

Degree of heterogeneity of attenuation The average interquartile range across constructs in Peterson's (1994) meta-analysis of Cronbach's alphas is approximately .16. If the standard deviation of the attenuation factors is .07, the interquartile range is approximately .16 for both levels of mean Cronbach's alpha. Therefore, when generating the attenuation factors, the standard deviation of the attenuation factor was .07. We did not examine the case of constant attenuation because it is unlikely to happen in real meta-analyses.

The range of Cronbach's alphas in Peterson's (1994) meta-analysis was .06 to .99. Therefore, we arbitrarily bounded the random attenuation factors from .224 to .995, which correspond to a range of .05 to .99 if Cronbach's alpha is used to compute the disattenuated correlation. Beta distribution was used to generate random attenuation factors with the desired range, mean, and standard deviation. To generate each random effect size, the population correlation and population attenuation factor were first computed, and then the population correlation was multiplied (attenuated) by the population attenuation factor to compute the actual population correlation used to generate the raw data, which were then used to compute the sample correlation.

Mean sample size Cheung and Chan (2008) examined two mean sample sizes, 100 and 300. We adopted these two levels and added the condition of mean sample size of 200. The case of 100 represents a reasonable minimum for the mean sample size in a meta-analysis that consists mainly of correlations, whereas the case of 300 simulates situations with small sampling variances. We added the level of 200 in order to have a better picture of the trend from 100 to 300.

Number of studies in a meta-analysis We selected three numbers of studies in a meta-analysis for investigation. First, Dieckmann, Malle, and Bodner (2009) found that the median number of studies in the meta-analyses they reviewed was 64. Therefore, we selected 60 as the medium level of number of studies to be investigated, to represent the common situation. We also examined two other situations, 30 and 100 studies. Previous investigations suggested that the samplewise-adjusted procedures did not perform well when the number of studies was 12. Therefore, we examined the case of 30 studies. Moreover, although it is uncommon for meta-analyses to have more than 100 studies, this level would allow us to examine the large-sample performance of the procedures.

Following Cheung and Chan (2008), the sample size in each study was randomly generated, with a standard deviation equal to one-fifth of the mean (e.g., if the mean was 100, the standard deviation was 20). A beta distribution was used to

simulate the skewed distribution of the sample sizes usually found in meta-analysis. The minimum sample size was 50, and the maximum sample size was 1,000.

Mean degree of dependence We examined two degrees of dependence (ρ_{rr}), .30 and .70, which represent moderate and large correlations between effect sizes from the same study, respectively. We believe these two levels are representative of the common degrees of dependence within studies. To generate dependent correlations, we adopted the same repeated measures model used by Cheung and Chan (2004, 2008, please refer to the 2004 article for the technical details). In that model, a linear regression model is used to generate two variables with the desired population correlation. First, n standard normal random numbers x , the scores of the independent variable, and n standard normal random numbers e , the scores of the residuals, are generated. The dependent variable, y , is then computed by $y = rx + \sqrt{(1-r^2)}e$, such that the population correlation between x and y is equal to r , and y follows a standard normal distribution. For any two correlations, r_a and r_b , to be correlated, two parameters are used, R_{xx} and R_{ee} , such that x_a and x_b have a population correlation of R_{xx} , and e_a and e_b have a population correlation of R_{ee} . This simulates the case that the same relationship was measured twice, and the predictors and residuals in these two trials were correlated. Cheung and Chan (2004) demonstrated that, if R_{xx} and R_{ee} are equal, ρ_{rr} is equal to $R_{xx}R_{ee}$. Therefore, two pairs of mean R_{xx} and R_{ee} , .55/.55 and .84/.84, are used to simulate the two levels of dependence.

Variation of degrees of dependence The degree of variation in dependence was not our focus, though we believe heterogeneous dependence is a common situation. Therefore, we set the standard deviation of R_{xx} and R_{ee} to .05. We generated the random values of R_{xx} and R_{ee} by a beta distribution with the range from .00 to .95. The degree of dependence varied both between and within studies.

Mean number of effect sizes across studies Cheung and Chan (2008) examined the case in which the ratio of the number of studies to the number of effect sizes is 3:4. That is, the mean number of effect sizes per study was about 1.33. The number of effect sizes of each study was randomly determined, such that the results were not limited to a specific pattern of distribution of effect sizes, as in Cheung and Chan (2004). We modified the procedure such that even the number of effect sizes varied across replications. First, we used a Gamma distribution (shape = 2, scale = 0.5) to generate a random number for the i th study, g_i . We set these values of shape and scale such that the distribution was skewed but the kurtosis was not too large (about 3 when the mean number of effect sizes was 3), reducing the likelihood of studies with extremely large

numbers of effect sizes. The mean of this random variable was 1. Second, the number of correlations for the i th study was $(f - 1)g_i + 1$, rounded to the nearest integer. When f was greater than or equal to 2, the mean number of correlations across studies was approximately f (only an approximation because the random numbers were rounded to the nearest integers). The distribution of effect sizes generated in this way simulated the realistic situation in which the distribution of number of effect sizes is usually skewed, with most studies having only a few effect sizes and some reporting larger numbers of effect sizes. In the present study, to cover the range of numbers of effect sizes per study usually found in meta-analytic review, we examined three levels of the mean number of effect sizes: 1.5, 2, and 3 (the corresponding f values were 1.57, 2, and 3).

Summary The total number of conditions was 3 levels of mean correlations $\times$ 3 degrees of heterogeneity $\times$ 2 levels of mean attenuation $\times$ 3 levels of mean sample sizes $\times$ 3 numbers of studies $\times$ 2 degrees of effect size dependence $\times$ 3 numbers of mean effect sizes, or 972. Each condition was analyzed by the six sample-wise procedures presented above.

Results

The patterns of results were similar for different degrees of mean population correlations (.10, .30, and .50) and different degrees of attenuation (attenuation factors .837 and .894). We also included only conditions with mean sample sizes of 100 and 300, because the results from these two conditions were sufficient to illustrate the trend of an outcome across our range of sample sizes. Thus, all the figures in the following section include only the other five factors: degree of heterogeneity, degree of dependence, mean sample size, number of studies, and mean number of effect sizes per study. The numerical results for all the conditions can be downloaded from https://osf.io/2q3w7/?view_only=3b8a2d7605764352b7e6608bb467507b.

Point estimates of the mean unattenuated population correlation

As is shown in Fig. 1, all procedures without correction for artifacts tended to underestimate the mean (unattenuated) population correlation. The number of studies, the mean sample size, the degree of dependence, the degree of heterogeneity, and the mean number of effect sizes all had little impact on the bias of these procedures. The two sample-wise procedures with correction for artifacts, on the other hand, show negligible biases in estimating the mean unattenuated population correlation. The number of studies, the degree of dependence, the degree of heterogeneity, and the mean number of effect sizes all had negligible impact on the bias of these procedures. These two procedures were slightly negatively biased when

the mean sample size was 100, but this bias was less than .005 and disappeared when the mean sample size was 300. This finding suggests that even when attenuation is present due to both artifacts and effect size dependency, the samplewise-adjusted procedure can still estimate the mean correlation with negligible bias.

Interval estimates of the mean unattenuated population correlations

The empirical coverage probabilities of the confidence intervals of the mean unattenuated correlations are shown in Fig. 2. For each sample-wise procedure without correction, we will interpret the associations between a factor and the coverage probability of a procedure only for conditions in which the biases shown Fig. 1 were small, or in which the biases decreased as the number of studies increased. If the bias of a procedure in a condition does not decrease as the number of studies increases, then the coverage probability of its confidence interval is expected to decrease as the number of studies increases. This is due to the fact that as the number of studies increases, the confidence interval becomes narrower and tends to move away from the true value. Therefore, if the bias of a procedure does not decrease with the number of studies, it is not worth examining how other factors may affect the coverage probabilities.

In light of the aforementioned reason, it is sufficient to note that the confidence intervals of all four procedures without correction for artifacts had low coverage probabilities in all conditions (less than 50% in most conditions). The coverage probabilities can be as small as 20%, or even 10% when the number of studies was 100.

On the other hand, the two samplewise-adjusted procedures with correction for artifacts (*SAdjWt_C* for samplewise-adjusted weighted and *SAdjInd_C* for samplewise-adjusted individual, in the figure) yielded coverage probabilities close to the nominal value of 95% in all conditions. Moreover, all five factors—number of studies, mean sample size, degree of dependence, degree of heterogeneity, and mean number of effect sizes per study—had negligible associations, if any, with the coverage probability of these two procedures with correction for artifacts.

Point estimates of the degree of heterogeneity (τ)

Nonzero degree of heterogeneity The bias in estimating the degree of heterogeneity is shown in Fig. 3. We decided to present the bias in estimating standard deviation because that is in the same metric as the effect size (correlation, in our case), and hence is easy to interpret. In general, when the degree of heterogeneity was nonzero, the four procedures without correction for artifacts tended to underestimate the degree of heterogeneity of the unattenuated population correlations (except for the sample-wise- np procedure, which

Fig. 1 Bias in estimating the mean unattenuated population correlations

tended to overestimate in some situations). The number of studies had negligible associations with the bias, whereas the bias was smaller when the mean sample size was 300. Consistent with previous studies (Cheung & Chan, 2004), the estimates of the sample-wise- n procedure decreased as the degree of dependence decreased (so that the assumption of highly correlated effect sizes in the sample-wise- n procedure was violated). On the other hand, the estimates of the sample-wise- np procedure decreased as the degree of dependence increased (so that the assumption of independent effect sizes in the sample-wise- np procedure was violated). The degrees of bias for these four procedures had negligible associations with the degree of heterogeneity. Finally, for the sample-wise- n and sample-wise- np procedures without correction, the larger the mean number of effect sizes per study, the larger the bias became.

The bias of these procedures without correction can be explained by the sampling variance of the attenuated correlations. This sampling variance is smaller than that after each is multiplied by a correction factor (the inverse of the attenuation factor) that is equal to or greater than 1, resulting in a narrow range for the estimate of the degree of heterogeneity. The exception is for the sample-wise- np procedure, which tended to overestimate the degree of heterogeneity when the degree of dependence was large, the degree of heterogeneity was small ($\tau = .05$), and the mean sample size was 100. One possible reason for this result was that the positive bias from using np as the effective sample size was larger than the negative bias due to ignoring attenuation in these conditions. The sample-wise- np procedure, even if there is no attenuation, still tends to overestimate the degree of heterogeneity, because it treats the within-cluster effect sizes as independent when they

Fig. 2 Empirical coverage probabilities of the confidence intervals of the mean unattenuated population correlations

are not, underestimating the variance due to sampling error and, hence, overestimating the degree of heterogeneity.

The two samplewise-adjusted procedures with correction for artifacts, on the other hand, showed a consistent pattern in estimating the degree of heterogeneity. They tended to underestimate across all conditions when the mean sample size was small (100), with samplewise-adjusted individual (*SAdjInd_C* in the figure) being more negatively biased. Holding the mean sample size constant, the number of studies had little impact on bias in the range we examined (30 to 100). However, when the mean sample size was 300, these two procedures yielded negligible biases in most conditions and outperformed most or all other procedures without correction for artifacts. The larger the degree of dependence and the larger the degree of heterogeneity, the less biased these two procedures became. Finally, the biases of these two procedures had negligible associations with the mean number of effect sizes per study.

Zero degree of heterogeneity When the degree of heterogeneity was zero—that is, when the unattenuated population correlation was the same for all studies—all the procedures, with or without correction for artifacts, tended to overestimate the degree of heterogeneity. This was expected because, in the simulation and in practice, negative estimates of the degree of heterogeneity are truncated to zero. Therefore, the expected value of the estimates cannot be zero even if the population parameter is equal to zero. In this special case, the best procedure is the one that has the smallest positive bias.

In general, three of the four procedures without correction had similar or larger biases than the procedures with correction, except for the sample-wise-*n* procedure, which was the least biased when the mean sample size was 100. Moreover, the bias of this procedure decreased with the mean number of effect sizes per study. However, it should be noted that this procedure tended to have the largest negative bias for the same mean sample size

Sn: Samplewise-N; SAdjInd: SAdj.Individual; SAdjInd_C: SAdj.Individual Corrected; SAdjWt_C: SAdj.Weighted Corrected; SAdjWt: SAdj.Weighted; Snp: Samplewise-Np; k: Mean number of studies; n: Mean sample sizes; tau: Degree of heterogeneity (in SD); ke: Mean number of effect sizes per study; dep: Degree of dependence, see text.

Fig. 3 Bias in estimating the degree of heterogeneity (τ)

when the degree of heterogeneity was nonzero. This negative bias resulted in smaller bias for the zero-heterogeneity condition, due to truncation. The advantage of the sample-wise- n procedure disappeared as the mean sample size increased.

For all six procedures, the bias in estimation decreased as the mean sample size increased, while this bias had negligible associations with the number of studies. Moreover, the larger the degree of dependence, the larger the biases of these six procedures were found to be. Finally, the larger the mean effect size per study, the larger the bias of the sample-wise- np procedure and the smaller the bias (due to the smaller the estimate) of the sample-wise- n procedure were found to be, consistent with the association between the mean number of effect sizes per study and the biases of these two procedures when the degree of heterogeneity was nonzero.

Although all procedures were biased due to truncation when the true degree of heterogeneity was zero, the two

samplewise-adjusted procedures with correction for artifacts outperformed all procedures without correction for artifacts when the average sample size increased (except for the sample-wise- n procedure, which had smaller bias for the reason presented above).

Empirical coverage probability of the 95% confidence interval of the degree of heterogeneity (τ)

The empirical coverage probabilities of the 95% confidence intervals in estimating the degree of heterogeneity (τ) are reported in Fig. 4. We selected certain patterns of results to interpret using the same logic adopted in interpreting the coverage probabilities of the confidence intervals when estimating the mean unattenuated correlation. When the degree of heterogeneity was nonzero, all four sample-wise procedures without correction for artifacts had suboptimal coverage

Fig. 4 Empirical coverage probabilities of the 95% confidence interval of the degree of heterogeneity (τ)

probabilities, though to varying degrees. Moreover, the coverage probability generally decreased with the number of studies, for the same reason discussed above in the case of interval estimation of the mean unattenuated correlations.

When the degree of heterogeneity was nonzero, the two samplewise-adjusted procedures with correction for artifacts had coverage probabilities close to the nominal value in all conditions. Again, the five factors—number of studies, mean sample size, degree of dependence, degree of heterogeneity, and mean number of effect sizes per study—had negligible relationships with the coverage probabilities.

When the population effect sizes were homogeneous, all six methods, including the samplewise-adjusted procedures, had suboptimal coverage probabilities. Moreover, though to varying degrees, the coverage probabilities decreased as the number of studies increased.

Taking into account the biases of these methods when the degree of heterogeneity was zero (see Fig. 3), this association can be explained by the decrease in sampling variance coupled with nonzero bias.

Discussion

The present study extended the samplewise-adjusted procedures to meta-analyzing dependent correlations corrected for artifacts. A simulation study was conducted in order to examine their performance in estimating the population mean unattenuated correlation and its true variance (the degree of heterogeneity). In estimating the population mean unattenuated correlation, the samplewise-adjusted procedures with correction for

artifact had negligible biases in point estimation, whereas all other procedures without correction substantially underestimated the mean unattenuated correlation. In the case of no attenuation due to artifacts, and thus no need for correction, Cheung and Chan (2004, 2008) found that the samplewise-adjusted procedures had negligible biases in estimating the mean population correlations in the presence of dependent correlations. Our findings revealed that, even with attenuation due to artifacts in the dependent correlations, the samplewise-adjusted procedures, modified in the present article, can also estimate unbiasedly the mean unattenuated correlation in the population.

In interval estimation, the samplewise-adjusted procedures with correction for artifacts yielded coverage probabilities close to the nominal value in nearly all conditions examined, whereas the confidence intervals of procedures without correction tended to miss the mean unattenuated correlation too often, with coverage probabilities lower than 50% in most conditions. Cheung and Chan (2008) found that the confidence intervals of the samplewise-adjusted procedures had coverage probabilities close to the nominal values in the presence of random within-study degree of dependence. Our results revealed that, even with between- and within-study random variation of attenuation due to artifacts, the modified samplewise-adjusted procedures can still yield confidence intervals with optimal coverage probabilities.

In estimating the degree of heterogeneity, when the degree of heterogeneity was nonzero, both samplewise-adjusted procedures with correction for artifacts outperformed the procedures without correction in point estimation and interval estimation. The samplewise-adjusted procedures with correction for artifacts tended to have the smallest biases and empirical coverage probabilities for the confidence intervals that were closest to the nominal value (95%). Cheung and Chan (2014) conducted a small-scale Monte Carlo study to examine the point and interval estimates of the samplewise-adjusted procedures in the absence of attenuation when $\tau = .10$ and found that these procedures had negligible biases and optimal coverage probabilities for the confidence interval. Our results show that the revised samplewise-adjusted procedures can also yield point estimates that are unbiased and confidence intervals with optimal coverage probabilities even in the presence of attenuation due to artifacts.

There is one caveat to our findings. When the degree of heterogeneity was zero, all six methods were biased in their point estimations, and the confidence intervals had low coverage probabilities. To our knowledge, these conditions had not been examined in previous studies. Cheung and

Chan (2004, 2008) used the Hunter–Schmidt approach, whereas we used the Hedges–Vevea approach. Cheung and Chan (2014) used the same approach that we did but did not investigate the case of homogeneous population correlations. Our finding in this condition may not be an issue in practice, because it is unusual for a meta-analysis with 30, or even 100, studies to have exactly the same unattenuated population correlations. Nevertheless, it is unclear whether this finding is due to the application of the Hedges–Vevea approach to correlations when the degree of heterogeneity is zero, or is a phenomenon specific to using the sample-wise procedures. Viechtbauer (2007) studied the performance of the Q-profile confidence interval used in our study and found it to have accurate coverage probabilities even when the degree of heterogeneity was zero. However, log odds ratio was the effect size measure used in that study. It is not clear whether those results can be generalized to meta-analyzing correlations. Further studies will be required in order to investigate this issue.

Limitations

There are several limitations to the revised samplewise-adjusted procedures and the Monte Carlo study. First, the samplewise-adjusted procedures require the attenuation factors for all correlations. It is not unusual to find primary studies without information on artifacts, such as reliability coefficient and construct validity, for some correlations. The method of artifact distribution adopted by Schmidt and Hunter (2015) can handle missing data in artifacts. In practice, imputation by means may be used as a tentative method. However, further efforts will be needed to extend the samplewise-adjusted procedures to allow for missing artifacts.

Second, we examined only sample-wise procedures because they can be extended to correct for artifacts. However, other methods can be used to meta-analyze dependent correlations in other situations, such as the three-level modeling approach or the robust variance estimates, as we presented in the brief review in the introduction. Further studies will be required in order to develop procedures to correct for artifacts in these approaches.

Finally, our study investigated only procedures to correct for artifacts in dependent correlations. Although, in principle, the original samplewise-adjusted procedure can be extended to other types of effect size measures, such as standardized mean differences, no studies have been conducted to examine this possibility. Correction for artifacts should also be a concern for other types of effect sizes. Further studies should be conducted to develop procedures to meta-analyze other types of dependent effect sizes with correction for artifacts.

Conclusion

To the best of our knowledge, there has been no meta-analytic procedure to correct for artifacts in dependent effect sizes. In the present study, we proposed an extension of the samplewise-adjusted procedures with correction for artifacts and examined a wide range of sample sizes, numbers of studies, mean population correlations, and degrees of attenuation. We also randomly generated the pattern of grouping of dependent correlations, such that our results are not restricted to a few specific patterns of grouping (e.g., each study with only one or two correlations, or a few studies with many correlations). Our findings suggest that the samplewise-adjusted procedures with correction for artifacts are viable alternatives to use when meta-analyzing dependent correlations that also vary in artifacts such as unreliability. The findings also reveal that, if attenuation due to artifacts is present but ignored, the point and interval estimates of both mean effect size and true variation (of the unattenuated mean effect size) can be adversely affected.

Author note This research was supported by a Multi-Year Research Grant (MYRG2014-00114-FSS) from the University of Macau. The simulation was performed in part at the High Performance Computing Cluster (HPCC), which is supported by the Information and Communication Technology Office (ICTO) of the University of Macau.

Appendix: Correction for artifacts

Suppose that the i th study contributes p_i correlations, r_{ij} , where $j = 1$ to p_i . Let b_{ijl} be the correlation between r_{ij} and r_{il} from the i th study. For the j th correlation of this study, the attenuation factor is A_{ij} . The j th disattenuated correlation, r_{Cij} , is r_{ij}/A_{ij} . Denote the sampling variance of the j th uncorrected correlation by σ_{ej}^2 ; thus, the sampling variance of r_{Cij} , denoted by v_{ij} , is σ_{ej}^2/A_{ij}^2 . To compute the within-study mean disattenuated correlation, r_{Ci} , we need to weight each correlation by its sampling variance. The weight of the j th disattenuated correlation is given by

$$w_{ij} = v_{ij}^{-1} = A_{ij}^2/\sigma_{ej}^2. \quad (\text{A1})$$

The within-study mean disattenuated correlation is then given by

$$r_{Ci} = \sum_{j=1}^{p_i} w_{ij} r_{Cij} / \sum_{j=1}^{p_i} w_{ij}. \quad (\text{A2})$$

Let the correlation between the *uncorrected* correlations r_{ij} and r_{il} be b_{ijl} ; in that case, the sampling variance of r_{Ci} is given by

$$\begin{aligned} \text{Var}[r_{Ci}] &= \left(\sum_{j=1}^{p_i} w_{ij} \right)^{-2} \text{Var} \left[\sum_{j=1}^{p_i} w_{ij} r_{Cij} \right] \\ &= \left(\sum_{j=1}^{p_i} w_{ij} \right)^{-2} \left[\sum_{j=1}^{p_i} w_{ij}^2 v_{ij} + \sum_{j=1}^{p_i} \sum_{l=1, l \neq j}^{p_i} w_{ij} w_{il} \sqrt{v_{ij} v_{il}} b_{ijl} \right] \\ &= \left(\sum_{j=1}^{p_i} w_{ij} \right)^{-2} \left[\sum_{j=1}^{p_i} w_{ij}^2 w_{ij}^{-1} + \sum_{j=1}^{p_i} \sum_{l=1, l \neq j}^{p_i} w_{ij} w_{il} \sqrt{w_{ij}^{-1} w_{il}^{-1}} b_{ijl} \right] \quad (\text{A3}) \\ &= \left(\sum_{j=1}^{p_i} w_{ij} \right)^{-2} \left[\sum_{j=1}^{p_i} w_{ij} + \sum_{j=1}^{p_i} \sum_{l=1, l \neq j}^{p_i} \sqrt{w_{ij} w_{il}} b_{ijl} \right] \\ &= \left(\sum_{j=1}^{p_i} \frac{A_{ij}^2}{\sigma_{ej}^2} \right)^{-2} \left[\sum_{j=1}^{p_i} \frac{A_{ij}^2}{\sigma_{ej}^2} + \sum_{j=1}^{p_i} \sum_{l=1, l \neq j}^{p_i} \frac{A_{ij} A_{il}}{\sqrt{\sigma_{ej}^2 \sigma_{el}^2}} b_{ijl} \right] \end{aligned}$$

As in the original sample-wise procedure (Cheung & Chan, 2004, 2008), the sampling variance of each *uncorrected* correlation is estimated by $\sigma_{ei}^2 = (1 - \bar{r}_i^2)^2 / (n_i - 1)$, instead of being estimated individually, where $\bar{r}_i$ is the weighted average *uncorrected* correlation for the i th study. Equation A2 then becomes

$$\begin{aligned} r_{Ci} &= \sum_{j=1}^{p_i} \left(A_{ij}^2 / \sigma_{ei}^2 \right) r_{Cij} / \sum_{j=1}^{p_i} \left(A_{ij}^2 / \sigma_{ei}^2 \right) \\ &= \sum_{j=1}^{p_i} A_{ij}^2 r_{Cij} / \sum_{j=1}^{p_i} A_{ij}^2 \end{aligned} \quad (\text{A4})$$

and Eq. A3 becomes

$$\begin{aligned} \text{Var}[r_{Ci}] &= \sigma_{ei}^4 \left(\sum_{j=1}^{p_i} A_{ij}^2 \right)^{-2} \left(\frac{1}{\sigma_{ei}^2} \sum_{j=1}^{p_i} A_{ij}^2 + \frac{1}{\sigma_{ei}^2} \sum_{j=1}^{p_i} \sum_{l=1, l \neq j}^{p_i} A_{ij} A_{il} b_{ijl} \right) \\ &= \sigma_{ei}^2 \left(\sum_{j=1}^{p_i} A_{ij}^2 \right)^{-2} \left(\sum_{j=1}^{p_i} A_{ij}^2 + \sum_{j=1}^{p_i} \sum_{l=1, l \neq j}^{p_i} A_{ij} A_{il} b_{ijl} \right) = \frac{\sigma_{ei}^2}{A_i^2} \end{aligned} \quad (\text{A5a})$$

$$A_i = \left(\sum_{j=1}^{p_i} A_{ij}^2 \right) / \sqrt{\sum_{j=1}^{p_i} A_{ij}^2 + \sum_{j=1}^{p_i} \sum_{l=1, l \neq j}^{p_i} A_{ij} A_{il} b_{ijl}} \quad (\text{A5b})$$

This notation is adopted to make it similar in form to the commonly used sampling variance estimate for disattenuated correlations, σ_e^2/A^2 . The square root of the denominator, A_i , can then be treated as the effective attenuation factor for the within-study mean disattenuated correlation that takes into account within-study dependence.

If all uncorrected correlations in the i th study are uncorrelated (i.e., b_{ijl} is zero for all j and l for the i th study), then $A_i = \sqrt{\sum_{j=1}^{p_i} A_{ij}^2}$. If all uncorrected correlations in the i th study are perfectly positively correlated (i.e., b_{ijl} is one for all j and l for the i th study), then $A_i = (\sum_{j=1}^{p_i} A_{ij}^2) / \sqrt{\sum_{j=1}^{p_i} A_{ij}^2 + \sum_{j=1}^{p_i} \sum_{l=1}^{p_i} 1 \neq j A_{ij} A_{il}}$. Assuming that $b_{ijl} = 0$, the procedure is equal to the sample-wise- np procedure (Cheung & Chan, 2014), in which the within-study mean is weighted by the product of the sample size and the number of effect sizes. Assuming that $b_{ijl} = 1$, the procedure is equal to the sample-wise- n procedure (Cheung & Chan, 2014), in which the within-study mean is weighted by the original sample size regardless of the number of effect sizes. These two possible values of A_i can then be treated as the boundaries for the effective attenuation factor.

If $A_{ij} = 1$ for all p_i correlations from the i th study, then $A_i = p_i / \sqrt{p_i + \sum_{j=1}^{p_i} \sum_{l=1}^{p_i} 1 \neq j b_{ijk}}$, and $1/A_i^2$ is the same correction factor developed by Cheung and Chan (2004, 2008) without correction for artifacts. In other words, the samplewise-adjusted procedure by Cheung and Chan can be considered a special case of the adjustment in Eq. A5, with no attenuation of the correlations.

In practice, the $p_i(p_i - 1)/2$ within-study inter-effect-size correlations (b_{ijk}) in the i th study are not known, as in the scenarios examined by Cheung and Chan (2004, 2008) without disattenuation. Let b_i be the unknown average of b_{ijk} in the i th study. The problem is how to estimate b_i or $\sum_{j=1}^{p_i} \sum_{l=1}^{p_i} 1 \neq j A_{ij} A_{jl} b_{ijl}$ in Eq. A5b. Two procedures to compute A_i and weigh the within-study means will be presented: the samplewise-adjusted individual procedure and the samplewise-adjusted weighted procedure.

The *samplewise-adjusted individual* procedure computes the effective attenuation factor for each group of dependent effect sizes separately. Because b_{ijl} are the correlations between the sample *uncorrected* correlations, b_i can be estimated without considering the correction coefficients A_{ijl} , as in the original samplewise-adjusted procedure:

$$\hat{b}_i = 1 - \frac{S_{ES_i}^2}{\sigma_{ei}^2}$$

After $\hat{b}_i$ is computed, it substitutes for b_{ijl} in to the correction factor formula:

$$A_i = \left(\sum_{j=1}^{p_i} A_{ij}^2 \right) / \sqrt{\sum_{j=1}^{p_i} A_{ij}^2 + \sum_{j=1}^{p_i} \sum_{l=1}^{p_i} 1 \neq j A_{ij} A_{il} \hat{b}_i} \quad (\text{A6})$$

The second procedure is the *samplewise-adjusted weighted* procedure. It uses a sample size weighted average of $\hat{b}_i$ across studies with dependent effect sizes in place of the individual $\hat{b}_i$. The effective attenuation factor is given by

$$A_i = \left(\sum_{j=1}^{p_i} A_{ij}^2 \right) / \sqrt{\sum_{j=1}^{p_i} A_{ij}^2 + \sum_{j=1}^{p_i} \sum_{l=1}^{p_i} 1 \neq j A_{ij} A_{il} \hat{b}} \quad (\text{A7})$$

where $\hat{b}$ is the weighted average of $\hat{b}_i$. This is the same samplewise-adjusted weighted procedure without correction for artifacts.

If σ_{ei}^2 is of the form $\lambda/(n_i - 1)$, where λ is a function of the sample effect size, such as $(1 - r^2)^2$ for the correlation coefficient, then the sampling variance of the within-study mean r_{Ci} can be computed like other correlations from single-effect-size studies, by using an effective sample size $n'_i = (n_i - 1)A_i^2 + 1$. When this effective sample size and the disattenuated correlation are used in the formula for the sampling variance of uncorrected correlations, $\sigma_e^2 = (1 - r^2)^2 / (n - 1)$, the resulting sampling variance is equal to the one based on the formula specifically for disattenuated correlations, $\sigma_e^2 = (1 - r^2)^2 / (n - 1)A^2$.

References

- Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational and Behavioral Statistics*, 17, 341–362. doi:<https://doi.org/10.3102/10769986017004341>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.
- Cheung, M. W. L. (2013). Multivariate meta-analysis as structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 20, 429–454. doi:<https://doi.org/10.1080/10705511.2013.797827>
- Cheung, M. W. L. (2015). *Meta-analysis: A structural equation modeling approach*. Chichester, UK: Wiley.
- Cheung, S. F., & Chan, D. K. S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology*, 89, 780–791. doi:<https://doi.org/10.1037/0021-9010.89.5.780>
- Cheung, S. F., & Chan, D. K. S. (2008). Dependent correlations in meta-analysis: The case of heterogeneous dependence. *Educational and Psychological Measurement*, 68, 760–777. doi:<https://doi.org/10.1177/0013164408315263>
- Cheung, S. F., & Chan, D. K. S. (2014). Meta-analyzing dependent correlations: An SPSS macro and an R script. *Behavior Research Methods*, 46, 331–345. doi:<https://doi.org/10.3758/s13428-013-0386-2>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

- Colquitt, J. A., Scott, B. A., Rodell, J. B., Long, D. M., Zapata, C. P., Conlon, D. E., & Wesson, M. J. (2013). Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives. *Journal of Applied Psychology*, 98, 199–236. doi:<https://doi.org/10.1037/a0031757>
- de Wit, F. R. C., Greer, L. L., & Jehn, K. A. (2012). The paradox of intragroup conflict: A meta-analysis. *Journal of Applied Psychology*, 97, 360–390. doi:<https://doi.org/10.1037/a0024844>
- Dieckmann, N. F., Malle, B. F., & Bodner, T. E. (2009). An empirical assessment of meta-analytic practice. *Review of General Psychology*, 13, 101–115.
- Gleser, L., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). New York, NY: Sage.
- Hedges, L. V., & Olkin, I. (1985). *Statistical models for meta-analysis*. New York, NY: Academic Press.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39–65. doi:<https://doi.org/10.1002/jrsm.5>
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486–504. doi:<https://doi.org/10.1037/1082-989X.3.4.486>
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.
- Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M. J., & Popovich, D. L. (2015). Toward a more nuanced understanding of the statistical properties of a median split. *Journal of Consumer Psychology*, 25, 652–665. doi:<https://doi.org/10.1016/j.jcps.2014.12.002>
- Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H. D., & O'Mara, A. (2009). Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research*, 79, 1290–1326. doi:<https://doi.org/10.3102/0034654309334143>
- Martinussen, M., & Bjørnstad, J. F. (1999). Meta-analysis calculations based on independent and nonindependent cases. *Educational and Psychological Measurement*, 59, 928–950. doi:<https://doi.org/10.1177/00131649921970260>
- Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, 21, 381–391. doi:<https://doi.org/10.1086/209405>
- R Core Team. (2018). *R: A language and environment for statistical computing* (Version 3.5.0) [Software environment]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org/
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Schmidt, F. L., Le, H., & Oh, I. (2009). Correcting for the distorting effects of study artifacts in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 317–333). New York, NY: Sage.
- Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62, 529. doi:<https://doi.org/10.1037//0021-9010.62.5.529>
- Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.
- Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, 4, 32–37. doi:<https://doi.org/10.1037/arc0000029>
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101. doi:<https://doi.org/10.1093/ije/dyq191>
- Tipton, E. (2013). Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods*, 4, 169–187. doi:<https://doi.org/10.1002/jrsm.1070>
- Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, 20, 375–393. doi:<https://doi.org/10.1037/met0000011>
- van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, 45, 576–594. doi:<https://doi.org/10.3758/s13428-012-0261-6>
- Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26, 37–52. doi:<https://doi.org/10.1002/sim.2514>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1–48. doi:<https://doi.org/10.18637/jss.v036.i03>
- Viswesvaran, C., Sanchez, J. I., & Fisher, J. (1999). The role of social support in the process of work stress: A meta-analysis. *Journal of Vocational Behavior*, 54, 314–334. doi:<https://doi.org/10.1006/jvbe.1998.1661>
- Wang, D., Waldman, D. A., & Zhang, Z. (2014). A meta-analysis of shared leadership and team effectiveness. *Journal of Applied Psychology*, 99, 181–198. doi:<https://doi.org/10.1037/a0034531>