Running head: DEPENDENT CORRELATIONS

Accepted for publication in *Educational and Psychological Measurement*:

Cheung, S. F., & Chan, D. K-S. (2008). Dependent correlations in meta-analysis: The case of heterogeneous interdependence. *Educational and Psychological Measurement, 68*, 760-777. doi:10.1177/0013164408315263

This is the accepted version and is not identical to the final published version. Please do not cite this version. The final published version can be obtained at: https://dx.doi.org/10.1177/0013164408315263

Dependent Correlations in Meta-Analysis: The Case of Heterogeneous Dependence

Shu Fai Cheung

University of Macau

Darius K-S. Chan

Chinese University of Hong Kong

Author Notes: Shu Fai Cheung, Department of Psychology; Darius K-S. Chan, Department of Psychology. Correspondence concerning this article should be addressed to Shu Fai Cheung, Department of Psychology, University of Macau, Av. Padre Tomás Pereira, Taipa, Macau, or by email (sfcheung@umac.mo). [Current email address: sfcheung@um.edu.mo]

Abstract

In meta-analysis, it is common to have dependent effect sizes, such as several effect sizes from the same sample but measured at different times. Cheung and Chan (2004) proposed the adjusted-individual and adjusted-weighted procedures to estimate the degree of dependence and incorporate this estimate in the meta-analysis. The present study extends the previous study by examining the case of heterogeneous degree of dependence. Simulation results reveal that these two procedures again generated less biased estimates of the degree of heterogeneity than the commonly used samplewise procedure, and were statistically more powerful to detect true variations. In addition, the adjusted-weighted procedure generated slightly less biased estimates of the degree of heterogeneity than the adjusted-individual weighted procedure across conditions. Future directions to further refine these procedures are discussed.

Keywords: Meta-analysis, dependent effect size

Dependent Correlations in Meta-Analysis: The Case of Heterogeneous Interdependence

Most popular statistical procedures in meta-analysis assume that the effect sizes are independent. However, it is not uncommon to encounter studies that contribute more than one effect size from same samples (e.g., Chapman, Uggerslev, Carrol, Piasentin, & Jones, 2005; Williams, McDaniel, & Nguyen, 2006; Zhao & Seibert, 2006). Statistical procedures have been proposed to handle this kind of dependent effect sizes (e.g., Gleser & Olkin, 1994; Hunter & Schmidt, 1990). Generalized least squares approach (Becker, 1992) and hierarchical linear modeling (Bryk & Raudenbush, 2002), when applied to meta-analysis, can also take into account the degree of dependence among the effect sizes. However, all these approaches require knowledge of the intercorrelations or covariances between the dependent effect sizes. In reality, these pieces of information are rarely available in published studies. This might be one of the reasons why the aforementioned analytic approaches were seldom used.

If the degree of dependence is unknown, one approach is to impute an estimate based on previous research. However, in most situations, it is difficult to decide what the best estimate should be. Another approach is to avoid the problem of non-independence by averaging the dependent effect sizes. This approach, denoted as the *samplewise procedure* by Cheung and Chan (2004), is a commonly adopted approach in published meta-analytic reviews, probably because of its simplicity and easiness to use. To avoid inflating the sample size, the original or average sample size is usually used as the weight for the average effect size of a particular sample or study. However, Cheung and Chan (2004) illustrated in their simulation study that the samplewise procedure tended to underestimate the degree of heterogeneity.

Cheung and Chan (2004) instead proposed to estimate the degree of dependence from the limited information and incorporate this estimate in the meta-analysis. Specifically,

these authors proposed two procedures, denoted as the *adjusted-individual procedure* and *adjusted-weighted procedure*, to incorporate the estimated degree of dependence in the meta-analysis. In their simulation study, they found that these two procedures were less biased than the samplewise procedure in estimating the degree of heterogeneity in meta-analysis. However, there were limitations in their study. For example, they only examined cases in which the degree of heterogeneity was homogeneous across samples - an unrealistic situation in meta-analysis. In the present study, we extend their study by examining a more realistic situation in actual meta-analysis, i.e., the case of heterogeneous degree of dependence across studies.

The Samplewise Procedure

In the present study, we focus our discussion on correlation (Pearson *r*), a popular estimate of effect sizes in meta-analysis of applied research. Suppose there are *K* studies in a meta-analysis, each contributes p_i correlations. We denote the *j*-th sample correlation contributed by the *i*-th study by r_{ij}, which is an estimate of the population correlation, ρ_i. We assume that the *K* studies are heterogeneous, and the population correlations follow an arbitrary distribution with a mean of $\rho_\bullet$ and a standard deviation of σ_ρ. To the extent that the p_i correlations in the *i*-th study are dependent, conventional meta-analytic procedures are not valid because the assumption of independence is violated. The samplewise procedure is then usually used to avoid the problem by computing an average correlation for the study with multiple effect sizes:

$$\bar{r}_i = \frac{\sum_{j=1}^{p_i} r_{ij}}{p_i} \qquad (1)$$

The *i*-th study is then treated as contributing one single correlation, $\bar{r}_i$, and any conventional meta-analytic procedure can be applied. In practice, the average sample size for the *i*-th study is usually used as the sample size associated with $\bar{r}_i$.

This samplewise procedure, although easy to use, tends to underestimate the degree of heterogeneity of the population correlation, σ_ρ (Cheung & Chan, 2004; Viswesvaran, Sanchez, & Fisher, 1999). In the extreme situation in which correlations are actually independent despite coming from the same sample, the average sample size will underestimate the sampling variance of $\bar{r}_i$. The samplewise procedure is appropriate only when the correlations are extremely dependent and hence the variation among dependent correlations is close to the sampling variance as estimated by the average sample size.

The Adjusted Procedures

Cheung and Chan (2004), based on the work of Martinussen and Bjornstad (1999), proposed two adjusted procedures that incorporate an estimate of the degree of dependence. As discussed above, one source of error in the samplewise procedure is the biased estimation of the sampling variance. If we can estimate the degree of dependence and adjust the sample size accordingly, then we can apply the common meta-analytic procedures on the averaged dependent correlations ($\bar{r}_i$). Using an approach similar to the estimate of interrater agreement by James, Demaree, and Wolf (1993), Cheung and Chan (2004) proposed an estimate of the degree of dependence in the *i*-th study, $\hat{\rho}_{rri}$, which is the estimated correlation between any two sample correlations contributed by the *i*-th study:

$$\hat{\rho}_{rri} \approx 1 - \frac{S_{ri}^2}{S_{ei}^2} \approx 1 - \frac{\sum\left(r_{ij} - \bar{r}_i\right)^2 / \left(p_i - 1\right)}{\left(1 - \bar{r}^{-2}\right)^2 / \left(n_i - 1\right)} \qquad (2)$$

This estimate is then used to compute the adjusted sample size for $\bar{r}_i$, $n_{*i} = \left(n_i - 1\right)/C_i + 1$, where

$$C_i = \frac{1+(p_i-1)\hat{\rho}_{rri}}{p_i}. \quad (3)$$

In the adjusted-individual procedure, $\hat{\rho}_{rr}$ is computed for each study with multiple effect sizes, and the sample size for each of these studies is adjusted by the estimated degree of dependence. In the adjusted-weighted procedure, the sample size weighted mean of $\hat{\rho}_{rr}$, denoted as $\overline{\hat{\rho}_{rr}}$, is used to adjust the sample sizes of all studies with multiple effect sizes.

These two procedures are preferable to the samplewise procedure. Without the need to know the correlations among all variables, the degree of dependence is estimated by the observed variance of the dependent correlations. Once the sample size has been adjusted by the estimated degree of dependence, similar to the samplewise procedure, any common procedure for meta-analysis can be applied.

Cheung and Chan (2004) demonstrated in a simulation study that the adjusted procedures resulted in fairly accurate estimate of the degree of heterogeneity when some of the correlations were dependent. However, as acknowledged by Cheung and Chan (2004), they limited their study to conditions in which the degree of dependence is constant across studies. That is, ρ_{rr} was constant both across and within all studies in the same replication. Although it seems to be acceptable as an initial investigation of a new procedure, this assumption is highly unrealistic in application. For example, suppose there are several longitudinal studies measuring the motivation-performance correlation at two different time points. It seems unreasonable to assume that the test-retest correlations (i.e., the degree of dependence) are identical across studies, even though the studies may differ in the time between the two time points, settings, samples, measures, jobs, and other variables. As another example, suppose a single study measured the attitude-intention correlation of several different behaviors (e.g., teeth brushing, blood donation, class attendance, etc.). It seems unreasonable to assume that the inter-correlations among the variables on different behaviors

are identical. Therefore, to have a better understanding of these adjusted procedures in realistic situations, we aim to investigate the case in which the degree of dependence, ρ_{rr}, varies both within and across studies.

The Case of Heterogeneous Degree of Dependence

First, we examine the sampling variance of the within sample average, $\bar{r}_i$. According to Cheung and Chan (2004, Appendix B, based on the derivation by Martinussen & Bjornstad, 1999), the sampling variance is given approximately by

$$V(\bar{r}_i) \approx \sigma_\rho^2 + \sigma_{ei}^2 \approx \sigma_\rho^2 + \frac{E\left[\left(1-\rho_i^2\right)^2\right]}{n_i - 1}\left[\frac{1+(p_i-1)\rho_{rri}}{p_i}\right], \quad (4)$$

where $E\left[\left(1-\rho_i^2\right)^2\right]$ is the expectation of $\left(1-\rho_i^2\right)^2$.

The derivation in Cheung and Chan (2004) does not assume homogeneous degree of dependence across samples. Therefore, we only examine analytically the case of heterogeneous degree of dependence within a sample. Let us assume that the degree of dependence within the *i*-th sample is not constant, and let b_{ist} be the correlation between r_{is} and r_{it}. then the equation becomes

$$V(\bar{r}_i) \approx \sigma_\rho^2 + \frac{E\left[\left(1-\rho_i^2\right)^2\right]}{n_i - 1}\left(\frac{p_i + \sum_{s=1}^{p_i}\sum_{t=1}^{p_i} b_{ist}}{p_i^2}\right), \text{ where } s \neq t. \quad (5)$$

Let $b_{i\bullet\bullet}$ be the average of b_{ist} for the *i*-th study, then equation becomes

$$V(\bar{r}_i) \approx \sigma_\rho^2 + \frac{E\left[\left(1-\rho_i^2\right)^2\right]}{n_i - 1}\left[\frac{1+(p_i-1)b_{i\bullet\bullet}}{p_i}\right]. \quad (6)$$

This formula is identical to Equation 4 for the case of homogeneous degree of dependence, except that ρ_{rri} is replaced by $b_{i\bullet\bullet}$.

Following the same rationale in Cheung and Chan (2004), this result suggests that we can adjust the sample size by C_i, where

$$C_i = \frac{1+(p_i-1)b_{i\bullet\bullet}}{p_i} \quad . \tag{7}$$

Therefore, if the average degree of dependence is known, the correction factor C_i and the adjusted sample size can be computed even when the degree of dependence is heterogeneous within the same study.

As apparent in the similarity between Equations 4 and 6, by substituting $b_{i\bullet\bullet}$ for ρ_{rri} in Equation B2 in Cheung and Chan (2004), it can be shown that we can use the same equation for $\hat{\rho}_{rri}$ as proposed by Cheung and Chan (2004) to estimate $b_{i\bullet\bullet}$:

$$\hat{b}_{i\bullet\bullet} \approx 1-\frac{S_{ri}^2}{S_{ei}^2} \approx 1-\frac{\sum\left(r_{ij}-\bar{r}_i\right)^2/\left(p_i-1\right)}{\left(1-\bar{r}^2\right)^2/\left(n_i-1\right)}. \tag{8}$$

In sum, although the derivation of the formulas are not exactly the same, it happens that the adjusted procedures proposed by Cheung and Chan (2004) are also applicable to the case of heterogeneous degree of dependence without any special treatment. Moreover, we made no assumption on the distribution of the degree of dependence within a sample, and did not require the knowledge of the standard deviation of the degree of dependence to estimate the correction factor.

The Monte Carlo Study

Although we can demonstrate that the adjusted procedures are applicable even when the degree of dependence varies within a study, a simulation study is necessary to empirically assess the performance of the procedures. We conducted a simulation study based on the one by Cheung and Chan (2004). In addition to introducing variations in the degree of dependence, we also made two major modifications to investigate the generalizability of their findings. First, instead of examining the two specific patterns used in their study, namely

minor multiplicity (each study contributes only one or two effect sizes) and skewed distribution (some studies contribute a large number of effect sizes, while most other studies contribute only one effect size), we randomly generated the pattern of distribution of the effect sizes across studies (details to be described in the Method section). Although this approach prevents us from investigating the effect of any specific pattern, this allows us to generalize our findings to all possible patterns.

Second, instead of using a normal distribution to randomly generate the population correlation (ρ_i) and the degrees of dependence (b_{ist}) for each study, we used the Beta distribution for this purpose. In all heterogeneous conditions, the population correlation and the degree of dependence were restricted to a minimum of zero and a maximum of .95 (we believe this is a reasonable upper limit for population correlation in most real studies). The parameters of the Beta distribution were determined such that the mean and standard deviation were of the desired values. For example, in the heterogeneous conditions with a average population correlation of .30 and a standard deviation of .10, the randomly generated correlation followed a Beta distribution with a mean of .30, a standard deviation of .10, a minimum of zero, and a maximum of .95. In other words, even though there was true variation across studies, the population correlation was never negative. The upper limit of .95 prevents any confounding in results due to possible severe deviation from normality of sampling distribution at extreme values of correlation. We believe this would not severely limit the coverage of our conditions because a population correlation of .95 or higher, though possible, is rare in actual studies. The use of a Beta distribution allows us to impose a lower bound and upper bound of the distribution while keeping the mean and standard deviation as specified, without the need to discard out-of-range values generated. In our cases, it is reasonable to impose the lower and upper bounds. For example, when the dependence is due to test-retest correlation, we may expect that the test-retest correlation varies across situation,

but we rarely expect this correlation to be negative in some cases and positive in other cases. Beta distribution allows us to specify the mean and standard deviation *a priori* while at the same time setting a lower and upper limits to the generated degree of dependence.

Power to Detect Genuine Heterogeneity

Meta-analyses of applied studies, especially those adopting the Hunter-Schmidt approach, emphasize on the estimation of the true variation and not on the heterogeneity significance test. This is natural because in the analysis of all the collected effect sizes, the total sample size and the number of studies are usually substantially large, leading to high statistical power to detect even a small true variation. However, most meta-analysts would also conduct subgroup analyses to investigate the variation of effect sizes in theoretically meaningful subgroups. The number of studies in each subgroup is usually much smaller for these subsequent analyses, some may be as small as ten or even fewer. In these subgroup analyses, it is not uncommon to find nonsignificant heterogeneity. Statistical power is a concern for this kind of subgroup analyses. To the extent that the samplewise procedure underestimates the true variation, the statistical power will also decrease. If the two adjusted procedures, after taking into account the estimated degree of dependence, lead to a more accurate estimation of the true variation, they should also have larger statistical power than the samplewise procedure. Therefore, in the present study, we also examined the empirical power of the three procedures in detecting true variation in population correlation.

Method

For each study, given the population correlation ρ and a sample size of n, n cases were generated based on the following formula:

$$y = \rho \times x + \left(\sqrt{1-\rho^2}\right) \times e, \tag{9}$$

where x and e are random variables following a normal distribution with a mean of zero and a standard deviation of one. The x's and y's then have an underlying standard normal

distribution with a population correlation equal to ρ. As detailed below, variation in population correlation was introduced by varying ρ across studies, and within-study dependence was introduced by forming common components for *x* and *e* for cases from the same study.

To maintain comparability, we manipulated factors similar to those used in Cheung and Chan (2004), with some modifications and improvements. Six factors were examined. First, the numbers of studies (K) examined were 12 and 60. Cheung and Chan did not find noticeable curvilinear relationship between the results and the number of studies. Moreover, the exact form of the relationship between the number of studies and the performance of the three procedures is not our focus. Therefore, we retained only these two conditions to represent the small and large scale meta-analyses respectively. The number of studies in actual meta-analysis varies widely. Moreover, even if the total number of studies is large in a meta-analysis, the number of studies can still be quite small for subgroup analysis in the same meta-analysis. Therefore, we selected the small and large conditions from Cheung and Chan to represent the two ends of the continuum of number of studies. The numbers of correlations (K_E) were 16 and 80 for conditions with 12 and 60 studies respectively. In other words, the $K:K_E$ ratio was fixed to 3:4. Instead of examining specific patterns of multiple effect sizes as in Cheung and Chan, we randomly generated the pattern for each replication. For example, if the number of studies was 12 and the number of effect sizes was 16, we first assigned one effect size to each of the 12 studies. For each of the remaining four effect sizes, we randomly assigned it to one of the 12 studies. This process was repeated until all the remaining four effect sizes were assigned. In other words, for each condition, some replications may have eight studies with one effect size and four studies with two effect sizes, while some replications may have 11 studies with one effect size and one study with five effect sizes. This procedure ensures that any possible pattern could have a chance to appear in the

simulation, thus allowing us to generalize the results to all possible patterns, instead of the two specific ones studied by Cheung and Chan (2004).

Second, the average sample sizes examined were 100 and 300. The average sample size in published meta-analysis varies widely, from around 100 (e.g., Dwight & Feigelson, 2000; Nesbit & Adesope, 2006;) to 300 or more (e.g., Gershoff, 2002; Thoresen, Kaplan, Barsky, Warren, & de Chermont, 2003; Wang, Jiao, Young, Brooks, & Olson, 2007). In the present study, we focused on meta-analyzing sample correlation. A sample size of 100 seems to be a reasonable minimum average sample size for this kind of studies. Therefore, an average sample size of 100 was selected as one of the conditions. To empirically examine the samplewise and the two adjusted procedures in meta-analysis with large average sample size, we selected an average sample size of 300 as the other condition. We extended the Cheung and Chan (2004) study by using a larger maximum average sample size because this is common in correlational studies to have sample sizes larger than 200. To generate the sample size for each sample in each replication, a method similar to that used by Field (2001) was adopted. The sample sizes across samples were drawn from a normal distribution with a standard deviation equal to the average sample size divided by five. For example, in the conditions that the average sample size equals 300, the sample sizes were drawn from a normal distribution with a standard deviation of 60.

Third, two different values of average population correlation, $\rho_{\bullet}$, were examined, namely .30 and .50, to represent the range of effect sizes commonly found in published meta-analysis. We focused on cases with population effect sizes vary across studies but in the same direction. Therefore, we did not include small average population (.10), which would not be able to obtain a true variation of .10 (in standard deviation) without leading to an extremely skewed distribution. Fourth, we manipulated the degree of heterogeneity. Three values of σ_{ρ}^2 were chosen, 0, .0025 and .01, which correspond to standard deviations of 0, .05,

and .10 respectively. Therefore, both fixed effects and random effects situations were examined (Hedges & Vevea, 1998). In the fixed effects situation, all studies have the same population correlation. In the random effects situations, the population correlation varies across studies. For the heterogeneous conditions (random effects situations) in which σ_ρ^2 is greater than zero, we defined the parameters of a beta distribution such that the mean was $\rho_{\bullet}$, the standard deviation was σ_ρ, the minimum was zero, and the maximum was .95.

The fifth and sixth factors were the average degree of dependence, $\rho_{\bullet rr}$, and standard deviation of the degree of dependence. We adopted the three degrees of dependence used in Cheung and Chan (2004), but extended their study by introducing random variation to the degree of dependence. Three levels of $\rho_{\bullet rr}$ were chosen, .09, .3025, and .49, to represent small, medium, and large effects as defined by Cohen (1988). Ideally, .10, .30, and .50 should be used. However, as discussed below, the parameters used were the square roots of these values, and for simplicity, we used values (.30, .55, and .70) to closely approximate the three levels of effect sizes. As in Cheung and Chan, a repeated measures model was adopted to model the dependence of the correlations. For example, the dependence might arise because the correlation between attitude and intention for the same behavior was assessed at two different times. To the extent that the attitude scores at the two different times were correlated, and the error terms (variation in intention not explained by attitude) were also correlated at the two different times, the correlation would be dependent. The parameters for the model are $\rho_{\bullet xx}$, the average correlation between *x*'s (attitude scores in the above example) within a sample, and $\rho_{\bullet ee}$, the average correlations between *e*'s (unexplained variance in intention in the above example) within a sample (for technical details, please see Cheung and Chan, 2004). The degree of dependence among the predictor variable *x* and among the error *e* could be translated into dependence among correlations, ρ_{rri}. As shown in Appendix C in Cheung

and Chan, when ρ_{xx} and ρ_{ee} both equal .30, degree of dependence is .09. To achieve the selected degrees of dependence, the configurations of $\rho_{\bullet xx}$ and $\rho_{\bullet ee}$ adopted were (.30, .30), (.55, .55), and (.70, .70). The corresponding average degrees of dependence are approximately .09, .3025, and .49 respectively. To manipulate the sixth factor, we adopted the beta distribution with a lower bound of zero and an upper bound of .95 to generate the ρ_{xx} and ρ_{ee} for each effect size. Two levels of standard deviations of ρ_{xx} and ρ_{ee}, .05 and .10, were used. This would result in the two degrees of variation in the degree of dependence within a study. We did not include the condition of constant degree of dependence because it has been studied in Cheung and Chan.

In sum, we examined six different factors in the Monte Carlo study, namely, the number of studies (12 and 16), the average sample size (100 and 300), the average population correlation (.30 and .50), the standard deviation of the population correlation (0, .05, and .10), the average degree of dependence (.09, 3025, and .49), and the degree of variation in the degree of dependence (standard deviation of ρ_{xx}/ρ_{ee}: 05 and .10). The total number of conditions was 72. The number of replications for each condition was 2,000. The theoretical distribution of the sample proportion, given a population proportion of .50 and 2,000 cases, is about .011. Therefore, 2,000 replications would lead to an accurate estimation of the statistical power.

Results

Estimating the Population Correlation ($\rho_{\bullet}$)

The mean estimated population correlation (or average population correlation, when between study heterogeneity was present) was close to the true value in all situations. For conditions with $\rho_{\bullet}$ =.30, the average estimated population correlation ranged from .297 to .301. For conditions with $\rho_{\bullet}$ =.50, the average estimated population correlation ranged from .496 to .500. When rounded to the second decimal place, the average estimated

population correlation was equal to the true value in all conditions. Therefore, the estimation of the average population correlation was practically unbiased in all situations and for all three methods.

Coverage Probabilities of the 95% and 90% Confidence Intervals

The empirical coverage probabilities of the 95% and 90% confidence intervals were examined. Due to similar results across some conditions, only the results for different degrees of heterogeneity (σ_ρ^2), different degrees of variation in degree of dependence (different standard deviations of ρ_{xx}/ρ_{ee}), and different degree of dependence ($\rho_{\bullet rr}$) were examined, with other conditions within the same combination of these three factors were collapsed together. The coverage probabilities are presented in Table 2. In general, the coverage probabilities were close to the nominated values, although slight over–coverage was observed when the population correlations were homogeneous ($\sigma_\rho = 0$) and when the samplewise procedure was used.

Estimating the Average Degree of Dependence ($\rho_{\bullet rr}$)

The average mean squared errors of the estimated average degree of dependence are presented in Table 1. In general, the accuracy of the two adjusted procedures was similar across conditions. The results were also similar for the two average sample sizes (100 and 300) and the two average population correlations (.30 and .50). The average estimates were closer to the population average degree of dependence when the number of studies was 60 than when it was 12. Moreover, the higher the average degree of dependence, the smaller the error was. In sum, the error was largest when the number of studies was 12 and the population average degree of dependence was small (.09).

Estimating the Degree of Heterogeneity (σ_ρ^2)

The means of the estimated degree of heterogeneity for σ_ρ^2 equal to .0000 and .0100 were reported in Figures 1 and 2 respectively. Examining these two conditions with the

intermediate condition (σ_ρ^2 equal to .0025) did not suggest any U-shaped relationship, and so results for σ_ρ^2 equal to .0025. Similarly, examining the three conditions of average degrees of dependence ($\rho_{\bullet rr}$ = .49, .30, and .09) did not suggest any U-shaped relationship. Therefore, only results for $\rho_{\bullet rr}$ equal to .49 and .09 were reported in each figure to ensure the readability of the figures, As shown in the figures, the samplewise procedure was the most negatively biased (i.e., underestimating the true variations) in most conditions, even when the average sample size was 300. The adjusted-weighted procedure, on the other hand, was the least biased in most conditions. The adjusted-individual procedure, though more biased than the adjusted-weighted procedure, still performed better than the samplewise procedure in most conditions.

In general, the larger the average sample size, the less biased the estimation for all three procedures. The benefit of large sample size was especially apparent for the samplewise procedure. On the other hand, as shown in Figure 2, increase in the number of studies also seemed to decrease the bias when true variations existed ($\sigma_\rho^2 > 0$). In most conditions, the bias when the average population correlation ($\rho_\bullet$) was medium (.50) was also smaller than when the average population correlation was small (.30).

When comparing conditions with different average degrees of dependence ($\rho_{\bullet rr}$), it was found that the bias of the samplewise estimate decreased when the average degree of dependence increased. That is, in general, the higher the degree of dependence (the thicker the line), the smaller the bias (the higher the line). The two adjusted procedures seemed to be similarly affected by the average degree of dependence, but to a lesser extent.

When considering the variation in the degree of dependence (standard deviation of ρ_{xx}/ρ_{ee}: 05 versus .10), the biases of both the samplewise procedure and the adjusted-individual procedure seemed to be smaller when the variation was larger. The

variation in the degree of dependence, on the other hand, had no consistent effect on the adjusted-weighted procedure.

Empirical Power to Detect True Variations

Due to space limit, only the results for K=12 were reported, because it is unlikely that the number of studies in subgroup analysis is as large as 60. In most conditions, the empirical power of the samplewise procedure was smaller than the two adjusted procedures, especially when the average sample size was 100. Averaged across conditions, empirical power of the samplewise procedure was 5.2% less than that of the adjusted-individual procedure, and 6.5% less than that of the adjusted-weighted procedure. The two adjusted procedures, on the other hand, had similar empirical power in most situations (average difference across situation was 1.4%). When the true variation was small (σ_ρ = .05), the power disadvantage of the samplewise procedure was as large as .10 even when the average sample was 300, that is, 10% less likely to detect the true variation when compared to the two adjusted procedures. The three procedures had similar empirical power when the average sample size was 300 and the true variation was large (σ_ρ = .10), ranged from 95% to nearly 100%. Last, the degree of dependence had negligible influence on the empirical power of each procedure.

Discussion

The simulation results provide further support for the two adjusted procedures. Even when the degree of dependence is heterogeneous both across and within studies, our results suggest that the adjusted-weighted and adjusted-individual procedures still yielded more accurate estimates of the degree of heterogeneity than the commonly used samplewise procedure. The two adjusted procedures were also statistically more powerful to detect true variation, especially when the number of studies was small. The present study also suggests that the advantage of the adjusted procedures over the samplewise procedures was not

specific to the two patterns of distribution of effect sizes across studies that were examined in Cheung and Chan (2004). In the present study, the patterns were randomly generated and so the results found could be in principle generalized to all possible patterns. We also examined the conditions with large average sample size (300), and found that the samplewise procedure was still the most biased procedure even when the average sample size was large. It should be noted that the large sample size did help to decrease its bias. In sum, our results suggest that the two adjusted procedures should be used instead of the samplewise procedure when some studies contribute more than one effect size and there is insufficient information to apply the analytic procedures available.

As mentioned earlier, the samplewise procedure essentially assumes that the several effect sizes given by the same study are highly dependent. Therefore, the bias of the samplewise procedure should decrease as the average degree of dependence increase. This relationship was found in Cheung and Chan (2004), and is replicated in the present study. Moreover, when comparing the two adjusted procedures, our results suggest that even when the degree of dependence varies across studies, one should use the weighted average estimated degree of dependence to adjust the sample size of all within study average correlations (the adjusted-weighted procedure), instead of, as would be suggested by intuition, adjusting each within study average using the estimated degree of dependence of that particular study (the adjusted-individual procedure). We believe that this finding may be related to the sampling error associated with the estimate of dependence for each study. The weighted average estimate of dependence should have much less sampling error than individual estimates. To the extent that the accuracy of the estimate of degree of heterogeneity depends on the expectation of the sampling error instead of the expectation of the estimated sampling error of each individual study, the weighted average instead of the individual estimates would produce a more accurate estimate of the degree of heterogeneity.

Note that this is the same rationale that Schmidt, Law, Hunter, and Rothstein (1993) used to support their use of average correlation, instead of individual correlations, to compute the sampling error of the weighted average correlation in meta-analysis.

Although we focused on the estimation and the significance test of the degree of heterogeneity, two aspects of the results also warranted discussion, the estimation of the average population correlation and the coverage probability of the confidence intervals. As in Cheung and Chan (2004), we found that the three procedures were accurate in estimating the average population correlation even when the degree of dependence varied within study. Moreover, the coverage probabilities of the 95% and 90% confidence intervals, though not exact, were close to the nominated level for all three procedures in most of the conditions examined. This suggests that if the main goal is to form point and interval estimates of the average population correlations, the three procedures would yield similar results.

The present study extends previous research by studying the more realistic situation of heterogeneous degree of dependence. However, four limitations should be noted. First, both the present study and the study by Cheung and Chan (2004) did not examine the presence of measurement error. To the best of our knowledge, the derivation of Hunter and Schmidt approach assumes independent effect sizes and may not work when the effect sizes are dependent, although there are some rules of thumb that are similar to the samplewise procedure. To expand the scope of application of both the adjusted procedures proposed and the Hunter and Schmidt approach, we need to incorporate the adjustment for dependent effect sizes into their approach with correction for artifacts. Second, both the present study and the study by Cheung and Chan (2004) were limited to the case that the within study population correlation is fixed. We believe this is a reasonable assumption, because differences in moderator variables, if any, are likely to be larger between studies than within studies. However, the scope of application of the adjusted procedures would be broader if this

assumption is relaxed. Future studies are necessary to revise the adjusted procedures to allow within study variation in the population correlation. Third, in the present study, we only investigated situations with positive average population correlations. It is possible that in some areas, a moderator may actually reverse the direction of an effect, leading to a small or even zero average population correlations. In this situation, the Type I error rates of the significance test of the zero average population correlations should be examined for the adjusted procedures. Fourth, like Cheung and Chan (2004), we adopted two assumptions that may not hold in some meta-analytic studies. We used a repeated measures model to generate within-study dependence. This is a plausible mechanism when the dependence is due to multiple assessments of the same or a similar pair of variables (e.g., attitude-intention correlation). However, there may be other forms of dependence in actual studies. Moreover, we assumed equal average ρ_{xx} and ρ_{ee}. Although we have relaxed the implausible assumption of equal ρ_{xx} and ρ_{ee} across studies made by Cheung and Chan (2004), it is reasonable to expect unequal average ρ_{xx} and ρ_{ee} in some actual studies. For example, for the case of repeated measures, the dependence among attitudes may be stronger than the dependence among measurement errors, leading to higher average ρ_{xx} than average ρ_{ee}. Although the derivation of the two proposed procedures does not depend on the mechanism leading to dependence, future studies should empirically examine the performance of the proposed procedure when these two assumptions are relaxed.

References

Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics, 17,* 341-362.

Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd Ed).* Thousand Oaks, CA: Sage.

Chapman, D. S., Uggerslev, K. L., Carrol, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant attraction to organizations and job choice: A meta-analytic review of the correlates of recruiting outcomes. *Journal of Applied Psychology, 90,* 928-944.

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.

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

Dwight, S. A., & Feigelson, M. E. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement, 60,* 340-360.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6,*161-180.

Gershoff, E. T. (2002). Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. *Psychological Bulletin, 128,* 539-579.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of research synthesis* (p. 399-355). New York: Russell Sage Foundation.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings (2nd Ed)*. Beverly Hills, CA: Sage.

James, L. R., Demaree, R. G., & Wolf, G. (1993). r_{wg}: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78*, 306-309.

Martinussen, M., & Bjornstad, J. F. (1999). Meta-analysis calculations based on independent and nonindependent cases. *Educational and Psychological Measurement*, 59, 928-950.

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research, 76*, 413-448.

Schmidt, F. L., Law, K., Hunter, J. E., & Rothstein, H. R. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78*, 3-12.

Thoresen, C. J., Kaplan, S. A., Barsky, A. P., Warren, C. R., & de Chermont, K. (2003). The affective underpinnings of job perceptions and attitudes: A meta-analytic review and integration. *Psychological Bulletin, 129*, 914-945.

Viswesvaran, C., Sanchez, J. L., & Fisher, J. (1999). The role of social support in the process of work stress: A meta-analysis. *Journal of Vocational Behavior, 54*, 314-334.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*, 219-238.

Williams, M. L., McDaniel, M. A., & Nguyen, N. T. (2006). A meta-analysis of the antecedents and consequences of pay level satisfaction. *Journal of Applied Psychology, 91*, 392-413.

Zhao, H., & Seibert, S. E. (2006). The Big Five personality dimensions and entrepreneurial status: A meta-analytical review. *Journal of Applied Psychology, 91*, 259-271.

Table 1. *The Average Mean Squared Error in Estimating the Average Degree of Dependence Across Conditions*

				Average Population Correlation ($\rho_{\bullet}$) = .30						Average Population Correlation ($\rho_{\bullet}$) = .50					
			Estimation	Individual			Weighted			Individual			Weighted		
			$\rho_{\bullet rI}$	.09	.30	.49	.09	.30	.49	.09	.30	.49	.09	.30	.49
σ_ρ^2	SD(ρ_{xx})	$\overline{N}$	K/K_E												
.0000	.05	100	12/16	.135	.107	.082	.136	.108	.083	.138	.108	.084	.139	.109	.084
			60/80	.040	.040	.028	.041	.040	.029	.039	.041	.029	.040	.042	.030
		300	12/16	.140	.108	.084	.143	.109	.085	.141	.109	.084	.144	.109	.084
			60/80	.039	.037	.027	.040	.038	.027	.040	.038	.027	.041	.039	.028
	.10	100	12/16	.140	.100	.087	.142	.100	.087	.141	.102	.086	.144	.102	.087
			60/80	.040	.039	.031	.041	.040	.032	.040	.040	.033	.041	.041	.033
		300	12/16	.138	.105	.086	.140	.105	.087	.138	.104	.087	.139	.104	.087
			60/80	.041	.039	.029	.042	.040	.030	.040	.040	.029	.041	.041	.029
.0025	.05	100	12/16	.139	.105	.082	.141	.106	.083	.137	.110	.087	.139	.110	.087
			60/80	.039	.040	.029	.040	.040	.030	.040	.040	.032	.042	.041	.033
		300	12/16	.137	.107	.085	.139	.108	.085	.137	.105	.083	.138	.105	.085
			60/80	.041	.040	.027	.042	.041	.029	.039	.040	.027	.040	.041	.028
	.10	100	12/16	.143	.104	.087	.145	.106	.087	.137	.105	.085	.138	.106	.085
			60/80	.043	.041	.028	.044	.041	.029	.038	.041	.031	.039	.042	.032
		300	12/16	.137	.104	.085	.138	.105	.086	.140	.107	.090	.142	.108	.090
			60/80	.040	.040	.028	.042	.040	.029	.040	.040	.028	.041	.041	.029
.0100	.05	100	12/16	.145	.107	.087	.148	.107	.088	.156	.113	.086	.159	.114	.086
			60/80	.039	.040	.030	.041	.040	.031	.042	.042	.031	.043	.042	.031
		300	12/16	.147	.105	.085	.149	.106	.087	.148	.110	.086	.150	.111	.086
			60/80	.043	.039	.027	.044	.040	.028	.042	.039	.028	.043	.040	.029
	.10	100	12/16	.144	.104	.087	.146	.105	.088	.151	.108	.092	.154	.109	.093
			60/80	.041	.040	.029	.042	.041	.030	.038	.040	.030	.039	.041	.031
		300	12/16	.145	.104	.087	.149	.106	.088	.151	.108	.086	.153	.109	.087
			60/80	.042	.039	.028	.043	.040	.029	.044	.040	.032	.045	.041	.033

Note: $\overline{N}$: Average sample sizes; K: Number of studies; K_E: Number of correlations; Individual: Adjusted-individual procedure; Weighted: Adjusted-weighted procedure; σ_ρ^2: Variation of population correlations; $\rho_{\bullet rI}$: Average degree of dependence.

Table 2. *The Coverage Probabilities of the 95% and 90% Confidence Intervals Across Conditions*

				Samplewise			Individual			Weighted		
	σ_ρ^2	SD(ρ_{xx})	$\rho_{\bullet rr}$	.09	.30	.49	.09	.30	.49	.09	.30	.49
95% Confidence Interval	.0000	.05		.97	.97	.97	.96	.96	.96	.96	.96	.96
		.10		.97	.96	.96	.96	.96	.96	.96	.96	.96
	.0025	.05		.94	.94	.94	.94	.94	.94	.94	.94	.94
		.10		.95	.94	.94	.94	.94	.94	.94	.94	.94
	.0100	.05		.94	.93	.94	.94	.93	.93	.94	.93	.93
		.10		.94	.94	.94	.93	.94	.94	.93	.94	.94
90% Confidence Interval	.0000	.05		.93	.93	.93	.92	.91	.91	.92	.92	.92
		.10		.93	.93	.92	.92	.91	.91	.92	.92	.91
	.0025	.05		.90	.89	.89	.89	.88	.89	.89	.88	.89
		.10		.89	.89	.90	.89	.89	.89	.89	.89	.89
	.0100	.05		.89	.88	.88	.88	.88	.88	.88	.88	.88
		.10		.88	.89	.89	.88	.89	.89	.88	.89	.89

Note: Samplewise: Samplewise procedure; Individual: Adjusted-individual procedure; Weighted: Adjusted-weighted procedure; σ_ρ^2: Variation of population correlations; $\rho_{\bullet rr}$: Average degree of dependence.

Figure Captions

Figure 1. The Average Estimated Degree of Heterogeneity (σ_ρ^2 = .0000).

Figure 2. The Average Estimated Degree of Heterogeneity (σ_ρ^2 = .0100).

TOP

Figure 1. The Average Estimated Degree of Heterogeneity (σ_ρ^2 = .0000)

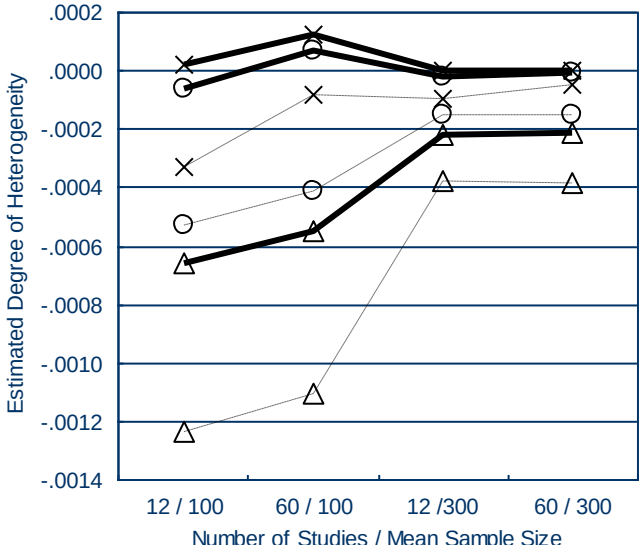

Figure 1a. Average ρ = .30, SD (ρ_{xx})=SD (ρ_{ee})=.05

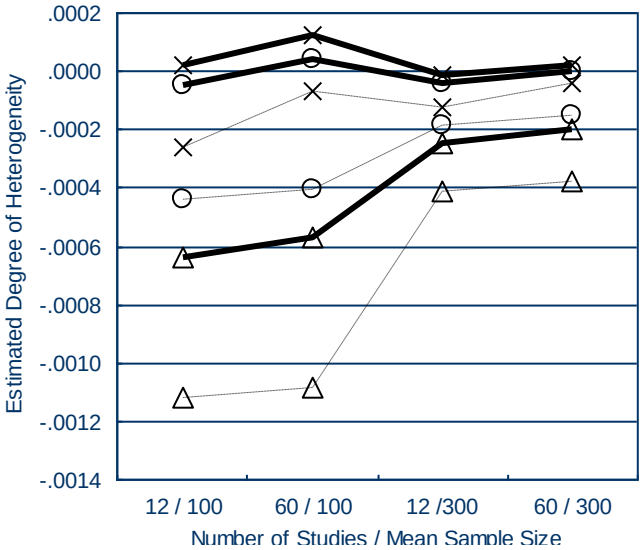

Figure 1b. Average ρ = .30, SD (ρ_{xx})=SD (ρ_{ee})=.10

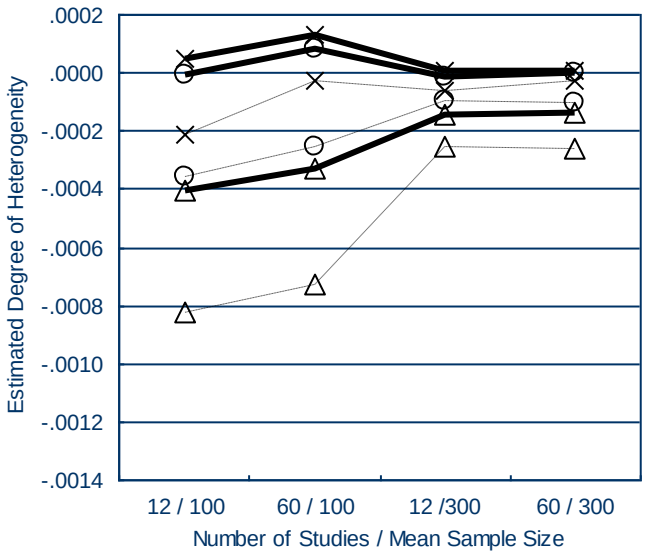

Figure 1c. Average ρ = .50, SD (ρ_{xx})=SD (ρ_{ee})=.05

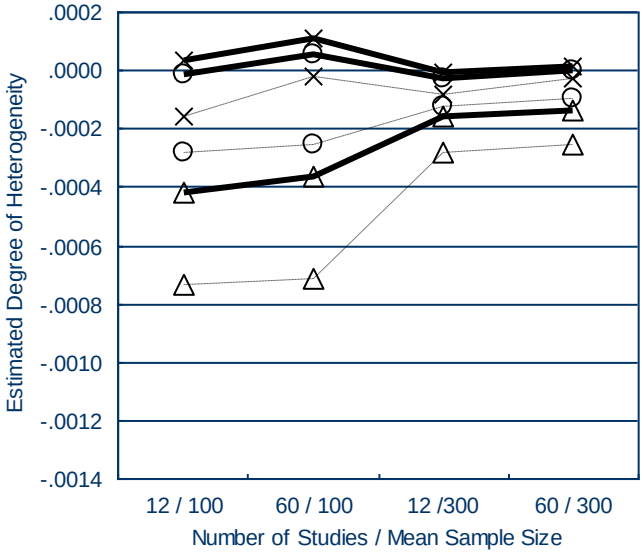

Figure 1d. Average ρ = .50, SD (ρ_{xx})=SD (ρ_{ee})=.10

Note: True population degree of heterogeneity (σ_ρ^2) is .0000. Average degree of dependence ($\rho_{\bullet rr}$) in parentheses (.49 and .09).

TOP

Figure 2. The Average Estimated Degree of Heterogeneity (σ_ρ^2 = .0100)

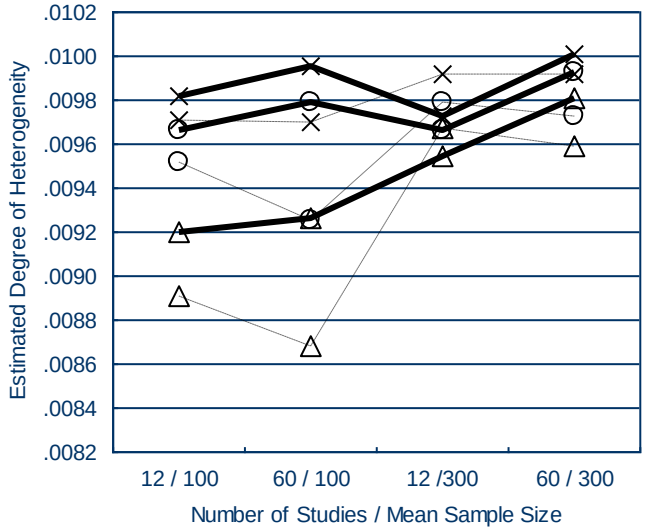

Figure 2a. Average ρ = .30, SD (ρ_{xx})=SD (ρ_{ee})=.05

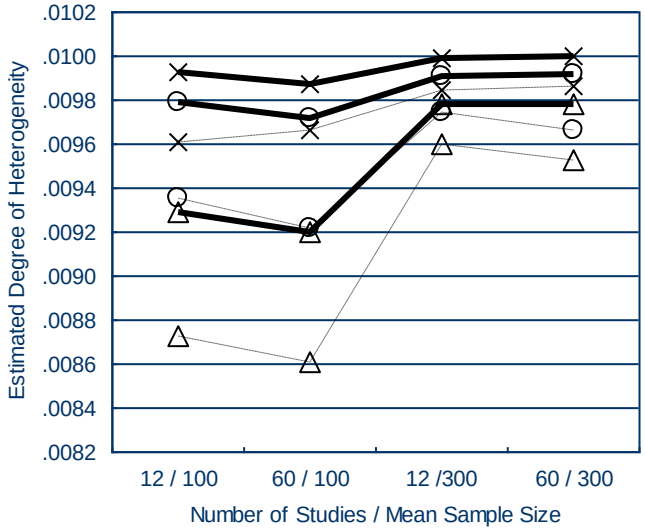

Figure 2b. Average ρ = .30, SD (ρ_{xx}) =SD (ρ_{ee})=.10

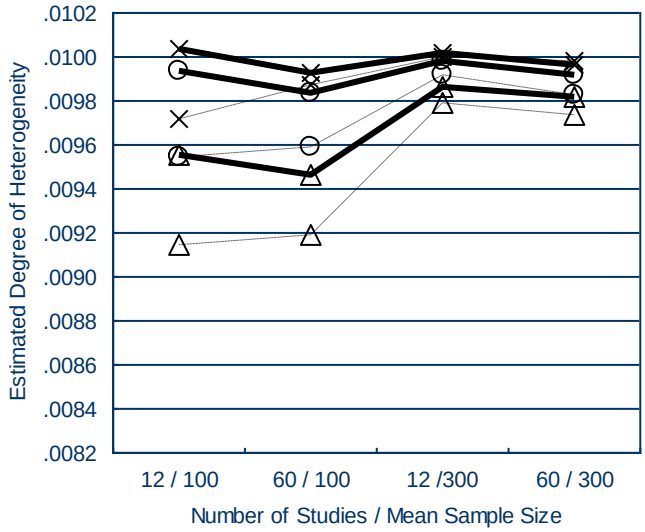

Figure 2c. Average ρ = .50, SD (ρ_{xx})=SD (ρ_{ee})=.05

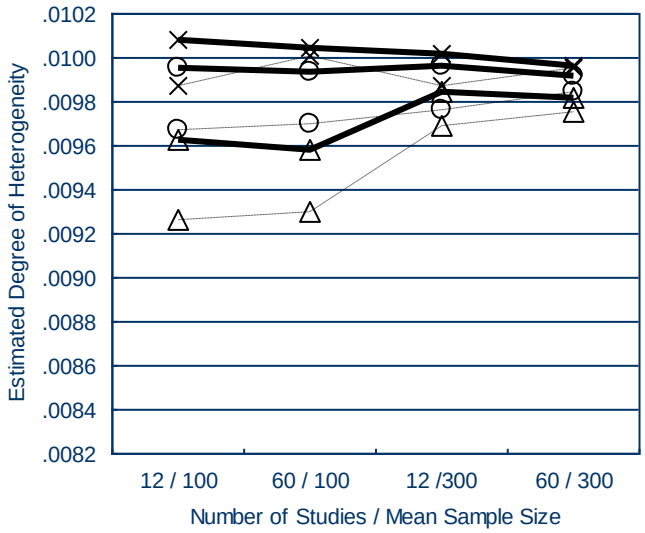

Figure 2d. Average ρ = .50, SD (ρ_{xx})=SD (ρ_{ee})=.10

Note: True population degree of heterogeneity (σ_ρ^2) is .0100. Average degree of dependence ($\rho_{\bullet rr}$) in parentheses (.49 and .09).