

# Sussex Research

## Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods.

Andy Field

### Publication date

01-06-2001

### Licence

This work is made available under the **Copyright not evaluated** licence and should only be used in accordance with that licence. For more information on the specific terms, consult the repository record for this item.

### Citation for this work (American Psychological Association 7th edition)

Field, A. (2001). *Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods*. (Version 1). University of Sussex. <https://hdl.handle.net/10779/uos.23310989.v1>

### Published in

Psychological Methods

### Link to external publisher version

<https://doi.org/10.1037/1082-989X.6.2.161>

### Copyright and reuse:

This work was downloaded from Sussex Research Open (SRO). This document is made available in line with publisher policy and may differ from the published version. Please cite the published version where possible. Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners unless otherwise stated. For more information on this work, SRO or to report an issue, you can contact the repository administrators at [sro@sussex.ac.uk](mailto:sro@sussex.ac.uk). Discover more of the University's research at <https://sussex.figshare.com/>

## Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods.

Article (Unspecified)

Field, Andy P (2001) Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, 6 (2). pp. 161-180. ISSN 1082-989X

This version is available from Sussex Research Online: <http://sro.sussex.ac.uk/id/eprint/713/>

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

### **Copyright and reuse:**

Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

# **META-ANALYSIS OF CORRELATION COEFFICIENTS: A MONTE CARLO COMPARISON OF FIXED- AND RANDOM-EFFECTS METHODS**

By

**Andy P. Field**

Dr. Andy P. Field  
Psychology Group  
School of Cognitive and Computing Science  
University of Sussex  
Falmer  
Brighton  
BN1 9QH  
UK

Tel: +44 (0)1273 877150  
Fax +44 (0)1273 671320  
Email [andyf@cogs.susx.ac.uk](mailto:andyf@cogs.susx.ac.uk)  
<http://www.cogs.susx.ac.uk/users/andyf>

Running Head: Meta-analysis of correlation coefficients

## **AUTHOR FOOTNOTE**

Correspondence concerning this article should be addressed to Andy P. Field, Psychology Group, School of Cognitive and Computing Science, University of Sussex, Falmer, Brighton, East Sussex, BN1 9QH, UK. Electronic mail may be sent to [andyf@cogs.susx.ac.uk](mailto:andyf@cogs.susx.ac.uk).

The author is indebted to four anonymous reviewers whose comments throughout the review process led to vast improvements in this work.

## **ABSTRACT**

The efficacy of the Hedges and colleagues', Rosenthal-Rubin, and Hunter-Schmidt methods for combining correlation coefficients were tested when population effect-sizes were both fixed and variable. After a brief tutorial on these meta-analytic methods the author presents two Monte Carlo simulations that compare these methods when the number of studies in the meta-analysis and the average sample size of studies were varied. In the fixed case the methods produced comparable estimates of the average effect-size; however, the Hunter-Schmidt method failed to control the Type I error rate for the associated significance tests. In the variable case, for both Hedges and colleagues' and Hunter-Schmidt methods: (1) Type I error rates were not controlled for meta-analyses including 15 or less studies; and (2) the probability of detecting small effects was less than 0.3. Some practical recommendations are made about the use of meta-analysis.

# **META-ANALYSIS OF CORRELATION COEFFICIENTS: A MONTE CARLO COMPARISON OF FIXED- AND RANDOM-EFFECTS METHODS**

## **INTRODUCTION**

Meta-analysis is a statistical technique by which information from independent studies is assimilated. Traditionally, social science literatures were assimilated through discursive reviews. However, such reviews are subjective and prone to ‘reviewer-biases’ such as the selective inclusion of studies, selective weighting of certain studies, and misrepresentation of findings (see Wolf, 1986). The inability of the human mind to provide accurate, unbiased, reliable and valid summaries of research (Glass, McGaw and Smith, 1981) created the need to develop more objective methods. Meta-analysis arguably provides the first step to such objectivity (see Schmidt, 1992), although it too relies on subjective judgements regarding study inclusion (and so is still problematic because of biased selections of studies, and the omission of unpublished data—the file drawer problem—see Rosenthal & Rubin, 1988). Since the seminal contributions of Glass (1976), Hedges and Olkin (1985), Rosenthal and Rubin (1978) and Hunter, Schmidt and Jackson (1982) there has been a meteoric increase in the use of meta-analysis. A quick search of a social science database<sup>1</sup> revealed over 2200 published articles using or discussing meta-analysis published between 1981 and 2000. Of these, over 1400 have been published since

1995 and over 400 in the past year. Clearly, the use of meta-analysis is still accelerating and consequently the question of which technique is best has arisen.

## **Methods of Meta-Analysis for Correlation Coefficients**

### Basic Principles

To summarise, an effect-size refers to the magnitude of effect observed in a study, be that the size of a relationship between variables or the degree of difference between group means. There are many different metrics that can be used to measure effect size: the Pearson product-moment correlation coefficient,  $r$ ; the effect-size index,  $d$ ; as well as odds ratios, risk rates, and risk differences. Of these, the correlation coefficient is used most often (Law, Schmidt & Hunter, 1994) and so is the focus of this study. Although various theorists have proposed variations on these metrics (for example, Glass's  $\Delta$ , Cohen's  $d$ , and Hedges's  $g$  are all estimates of  $\delta$ ), conceptually each metric represents the same thing: a standardized form of the size of the observed effect. Whether correlation coefficients or measures of differences are calculated is irrelevant because either metric can be converted into the other, and statistical analysis procedures for different metrics differ only in how the standard errors and bias corrections are calculated (Hedges, 1992).

In meta-analysis, the basic principle is to calculate effect sizes for individual studies, convert them to a common metric, and then combine them to obtain an average effect size. Studies in a meta-analysis are typically weighted by the accuracy of the effect size they provide (i.e. the sampling precision), which is achieved by using the sample size (or a function of it) as a weight. Once the mean effect size has

been calculated it can be expressed in terms of standard normal deviations (a  $Z$  score) by dividing by the standard error of the mean. A significance value (i.e. the probability,  $p$ , of obtaining a  $Z$  score of such magnitude by chance) can then be computed. Alternatively, the significance of the average effect size can be inferred from the boundaries of a confidence interval constructed around the mean effect size.

Johnson, Mullen and Salas (1995) point out that meta-analysis is typically used to address three general issues: central tendency, variability and prediction. Central tendency relates to the need to find the expected magnitude of effect across many studies (from which the population effect size can be inferred). This need is met by using some variation on the average effect size, the significance of this average or the confidence interval around the average. The issue of variability pertains to the difference between effect sizes across studies and is generally addressed using tests of the homogeneity of effect sizes. The question of prediction relates to the need to explain the variability in effect sizes across studies in terms of moderator variables. This issue is usually addressed by comparing study outcomes as a function of differences in characteristics that vary over all studies. As an example, differences in effect sizes could be moderated by the fact that some studies were carried out in the USA whereas others were conducted in the UK.

### Fixed versus Random Effects Models

So far, we have seen that meta-analysis is used as a way of trying to ascertain the true effect sizes (i.e. the effect sizes in a population) by combining effect sizes from individual studies. There are two ways to conceptualise this process: fixed effects and



random effects models<sup>2</sup>. Hedges (1992) and Hedges and Vevea (1998) explain the distinction between these models wonderfully. In essence, in the fixed effect conceptualisation, the effect sizes in the population are fixed but unknown constants. As such, the effect size in the population is assumed to be the same for all studies included in a meta-analysis (Hunter & Schmidt, in press). This situation is called the homogenous case. The alternative possibility is that the population effect sizes vary randomly from study to study. In this case each study in a meta-analysis comes from a population that is likely to have a different effect size to any other study in the meta-analysis. So, population effect sizes can be thought of as being sampled from a universe of possible effects — a ‘superpopulation’ (Hedges, 1992, Becker, 1996). This situation is called the heterogeneous case. To summarise, in the random effects model studies in the meta-analysis are assumed to be only a sample of all possible studies that could be done on a given topic whereas in the fixed effect model the studies in the meta-analysis are assumed to constitute the entire universe of studies (Hunter & Schmidt, in press).

In statistical terms the main difference between these models is in the calculation of standard errors associated with the combined effect size. Fixed effects models use only within-study variability in their error term because all other ‘unknowns’ in the model are assumed not to affect the effect size (see Hedges, 1992; Hedges & Vevea, 1998). However, in random effects models it is necessary to account for the errors associated with sampling from populations that themselves have been sampled from a superpopulation. As such the error term contains two components: within-study variability and variability arising from differences between studies (see Hedges &

Vevea, 1998). The result is that standard errors in the random-effects model are typically much larger than in the fixed case if effect sizes are heterogeneous and, therefore, significance tests of combined effects are more conservative.

In reality the random effects model is probably more realistic than the fixed effects model on the majority of occasions (especially when the researcher wishes to make general conclusions about the research domain as a whole and not restrict their findings to the studies included in the meta-analysis). Despite this fact, the National Research Council (1992) reports that fixed effects models are the rule rather than the exception. Osburn and Callender (1992) have also noted that real-world data are likely to have heterogeneous population effect sizes even in the absence of known moderator variables (see also Schmidt and Hunter, 1999). Despite these observations, Hunter and Schmidt (in press) reviewed the meta-analytic studies reported in *Psychological Bulletin* (a major review journal in psychology) and found 21 studies reporting fixed-effects meta-analyses but none using random effects models. Although fixed-effect models have attracted considerable attention (Hedges, 1992, 1994a,b), as Hedges and Vevea (1998) point out, the choice of model depends largely on the type of inferences that the researcher wishes to make: fixed-effect models are appropriate only for conditional inferences (i.e. inferences that extend only to the studies included in the meta-analysis) whereas random-effects models facilitate unconditional inferences (i.e. inferences that generalise beyond the studies included in the meta-analysis). For real-world data in the social sciences researchers typically wish to make unconditional inferences and so random-effects models are often more appropriate.

Over the last 20 years three methods of meta-analysis have remained popular (see Johnson, Mullen & Salas, 1995): the methods devised by Hedges, Olkin and colleagues, Rosenthal and Rubin (see Rosenthal, 1991), and Hunter and Schmidt (1990)<sup>3</sup>. Hedges and colleagues (Hedges & Olkin, 1985; Hedges, 1992; Hedges & Vevea, 1998) have developed both fixed- and random-effects models for combining effect sizes, whereas Rosenthal (1991) presents only a fixed-effects model, and Hunter and Schmidt present what they have labelled a random-effects model (see Schmidt & Hunter, 1999). Although Johnson et al. (1995) overview these three meta-analytic techniques, they did not use the methods for correlation advocated by Hedges and colleagues (or use the random-effects versions) and Schmidt and Hunter (1999) have made subsequent observations about the correct use of their method. Therefore, an overview of the techniques used in the current study, with reference to the original sources, is included as a pedagogical source for readers unfamiliar with meta-analysis of correlation coefficients.

#### Hedges-Olkin and Rosenthal-Rubin Method

For combining correlation coefficients, Hedges & Olkin (1985), Hedges and Vevea (1998) and Rosenthal and Rubin (see Rosenthal, 1991) are in agreement about the method used. However, there are two differences between the treatments that Hedges and colleagues and Rosenthal and Rubin have given to the meta-analysis of correlations. First, Rosenthal (1991) does not present a random effects version of the model. Second, to estimate the overall significance of the mean effect size, Rosenthal

and Rubin generally advocate that the probabilities of each effect size occurring by chance are combined (see Rosenthal, 1991; Rosenthal & Rubin, 1982).

### Fixed-Effects Model

When correlation coefficients are used as the effect-size measure, Hedges and Olkin and Rosenthal and Rubin both advocate converting these effect sizes into a standard normal metric (using Fisher's  $r$ -to- $Z$  transformation) and then calculating a weighted average of these transformed scores. Fisher's  $r$ -to- $Z$  transformation (and the conversion back to  $r$ ) is described in equation (1). The first step, therefore, is to use this equation to convert each correlation coefficient into its corresponding  $Z$  value (see Field, 1999 for an example).

$$z_{r_i} = \frac{1}{2} \text{Log}_e \left( \frac{1 + r_i}{1 - r_i} \right) \qquad r_i = \frac{e^{(2z_i)} - 1}{e^{(2z_i)} + 1} \qquad (1)$$

The transformed effect sizes are then used to calculate an average in which each effect size is weighted. Equation (2) shows that the transformed effect size of the  $i$ th study is weighted by a weight for that particular study ( $w_i$ ).

$$\bar{z}_r = \frac{\sum_{i=1}^k w_i z_{r_i}}{\sum_{i=1}^k w_i} \qquad (2)$$

Hedges and Vevea (1998) note that effect sizes based on large samples will be more precise than those based on small samples and so the weights should reflect the increased precision of large studies. In fact, the optimal weights that minimise the variance are the inverse variances of each study (see Hedges & Vevea, 1998, equation

2), and for correlation coefficients the individual variance is the inverse of the sample size minus three (see Hedges & Olkin, 1985, p. 227 and p. 231).

$$w_i = \frac{1}{v_i} \qquad v_i = \frac{1}{n_i - 3} \qquad \therefore w_i = n_i - 3$$

As such, the general equation for the average effect size given in equation (2) becomes equation (3) for correlation coefficients (this is equation 4.16 in Rosenthal, 1991, p. 74).

$$\bar{z}_r = \frac{\sum_{i=1}^k (n_i - 3) z_{r_i}}{\sum_{i=1}^k (n_i - 3)} \quad (3)$$

The sampling variance of this average effect size is simply the reciprocal of the sum of weights (Hedges and Vevea, 1998, equation 4) and the standard error of this average effect size is simply the square root of the sampling variance. So, in its general form the standard error is:

$$SE(\bar{z}_r) = \sqrt{\frac{1}{\sum_{i=1}^k w_i}} \quad (4)$$

Given that for correlation coefficients the weights are simply  $n - 3$ , the standard error becomes:

$$SE(\bar{z}_r) = \sqrt{\frac{1}{\sum_{i=1}^k (n_i - 3)}} \quad (5)$$

Hedges and colleagues recommend that a  $z$ -score of the mean effect size be calculated by simply dividing the mean effect size by its standard error (see equation

(6)). The probability of obtaining that value of  $\underline{Z}$  can then be calculated using the standard normal distribution (e.g. Field, 2000, p. 471). However, Rosenthal and Rubin recommend that the probability of obtaining the average effect size be calculated by combining the individual probability values of each correlation coefficient (see Rosenthal, 1991, p. 85-86, equation 4.31). This is the only respect in which the Rosenthal-Rubin and Hedges-Olkin fixed-effects methods differ.

$$Z = \frac{\bar{z}_r}{SE(\bar{z}_r)} \quad (6)$$

Finally, to test the homogeneity of effect sizes across studies, the squared difference between the observed transformed  $\underline{r}$  and the mean transformed  $\underline{r}$  is used. To create a chi-square statistic some account has to be taken of the variance of each study and as before, for correlation coefficients the variance is just the sample size minus 3 (see Hedges & Vevea, 1998, equation 7). This gives us the statistic  $\underline{Q}$  in Equation (7), which has a chi-square distribution (Rosenthal, 1991, equation 4.15, p. 74; Hedges & Olkin, 1985, equation 16, p. 235; Hedges & Vevea, 1998, equation 7, p. 490).

$$Q = \sum_{i=1}^k (n_i - 3) (z_{r_i} - \bar{z}_r)^2 \quad (7)$$

#### Random –effects model

Rosenthal (1991) does not present a random effects version of the model previously described. However, Hedges and Olkin (1985) and Hedges and Vevea (1998) clearly elaborate on how a random-effects model can be calculated. The main difference in the random effects model is that the weights are calculated using a

variance component that incorporated between-study variance in addition to the within-study variance used in the fixed-effect model. This between-study variance is denoted by  $\tau^2$  and is simply added to the within-study variance. As such the weights for the random-effects model ( $w_i^*$ ) are (see Hedges & Vevea, 1998, equation 13):

$$w_i^* = \frac{1}{v_i + \tau^2} \qquad v_i = \frac{1}{n_i - 3} \qquad w_i^* = \left( \frac{1}{n_i - 3} + \tau^2 \right)^{-1}$$

These new weights can simply be used in the same way as for the fixed-effects model to calculate the mean effect-size, its standard error and the  $z$ -score associated with it (by replacing the old weights with the new weights in equations 2, 4 and 6).

The question arises of how the between-study variance might best be estimated. Hedges and Vevea (1998) provide equations for estimating the between-study variance based on the weighted sum of squared errors,  $Q$  (see equation (7)), the number of studies in the meta-analysis,  $k$ , and a constant,  $c$  (see equation (9)).

$$\tau^2 = \frac{Q - (k-1)}{c} \tag{8}$$

The constant is calculated using the weights from the fixed effects model:

$$c = \sum_{i=1}^k w_i - \frac{\sum_{i=1}^k (w_i)^2}{\sum_{i=1}^k w_i} \tag{9}$$

When combining correlation coefficients the weights are just  $n - 3$  and the constant, therefore, becomes:

$$c = \sum_{i=1}^k (n_i - 3) - \frac{\sum_{i=1}^k (n_i - 3)^2}{\sum_{i=1}^k (n_i - 3)} \tag{10}$$

If, however, the estimate of between-study variance,  $\tau^2$ , yields a negative value then it is set at zero (because the variance between-studies cannot be negative).

Finally, the estimate of homogeneity of study effect sizes is calculated in the same way as for the fixed-effect model. In short, the only difference in the random-effects models is that the weights used to calculate the average and its associated standard error now include a between-study component that is estimated using equation (8).

### Hunter and Schmidt Method

Hunter and Schmidt advocate a single method (a random-effects method) based on their belief that fixed-effects models are inappropriate for real-world data and the type of inferences that researchers usually want to make (Hunter & Schmidt, in press)<sup>4</sup>. Hunter and Schmidt's method is thoroughly described by Hunter, Schmidt & Jackson (1982) and Hunter and Schmidt (1990). In its fullest form, this method emphasises the need to isolate and correct for sources of error such as sampling error and reliability of measurement variables. Although there is rarely enough information reported in a study to use the full Hunter and Schmidt technique, even in its simplest form it still differs from the method advocated by Hedges and colleagues and Rosenthal and Rubin. The main difference is in the use of untransformed effect-size estimates in calculating the weighted mean effect size. As such, central tendency is measured using the average correlation coefficient in which untransformed correlations are weighted by the sample size on which they are based. Equation (11) shows how the mean effect size is estimated and it differs from



equations (2) and (3) in that the weights used are simply the sample sizes on which each effect size is based, and each correlation coefficient is not transformed.

$$\bar{r} = \frac{\sum_{i=1}^k n_i r_i}{\sum_{i=1}^k n_i} \quad (11)$$

Like Hedges and colleagues' method, the significance of the mean effect size is obtained by calculating a  $\underline{Z}$  score by dividing the mean by its standard error. However, the estimate of the standard error is different in Hunter and Schmidt's method and there has been some confusion in the literature about how the standard error is calculated. Johnson et al. (1995) reported the equation of the variance across studies (the frequency weighted average squared error reported by Hunter and Schmidt 1990, p. 100). The square root of this value should then be used to estimate the standard deviation (as in Equation (12)). The best estimate of the standard error is to divide this standard deviation of the observed correlation coefficients by the square root of the number of studies being compared (Osburn & Callender, 1992; Schmidt et al., 1988). Therefore, as Schmidt and Hunter (1999) have subsequently noted, the equation of the standard deviation used by Johnson et al. should be further divided by the square root of the number of studies being assimilated. Equations (12) and (13) show the correct version (according to Schmidt & Hunter, 1999) of the standard deviation of the mean and the calculation of the standard error. The  $\underline{Z}$  score is calculated simply by dividing the mean effect size by the standard error of that mean (Equation (14)).

$$SD_r = \sqrt{\frac{\sum_{i=1}^k n_i (r_i - \bar{r})^2}{\sum_{i=1}^k n_i}} \quad (12)$$

$$SE_{\bar{r}} = \frac{SD_r}{\sqrt{k}} \quad (13)$$

$$Z = \frac{\bar{r}}{SE_{\bar{r}}} \quad (14)$$

In terms of homogeneity of effect sizes, again a chi-square statistic is calculated based on the sum of squared errors of the mean effect size (see p. 110-112 of Hunter and Schmidt, 1990). Equation (15) shows how the chi-square statistic is calculated from the sample size on which the correlation is based ( $n_i$ ), the squared errors between each effect size and the mean, and the variance.

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - 1)(r_i - \bar{r})^2}{(1 - \bar{r}^2)^2} \quad (15)$$

### Comparison of the Methods

There are two major differences between the methods described. The first difference is the use of transformed or untransformed correlation coefficients. The Fisher transformation is typically used to eliminate a slight bias in the untransformed correlation coefficient: the transformation corrects for a skew in the sampling distribution of  $\underline{r}$ s that occurs as the population value of  $\underline{r}$  becomes further from zero (see Fisher, 1928). Despite the theoretical basis for this transformation Hunter and Schmidt (1990) have long advocated the use of untransformed correlation coefficients using theoretical arguments to demonstrate biases arising from Fisher's

transformation (see Hunter, Schmidt & Coggin, 1996). Hunter and Schmidt (1990) note that 'the Fisher  $Z$  replaces a small underestimation or negative bias by a typically small overestimation, or positive bias, a bias that is always greater in absolute value than the bias in the untransformed correlation' (p. 102, see also Hunter et al., 1996; Schmidt, Gast-Rosenberg and Hunter, 1980; Schmidt, Hunter & Raju, 1988; Field, 1999).

Some empirical evidence does suggest that transforming the correlation coefficient can be beneficial. Silver and Dunlap (1987) claimed that meta-analysis based on Fisher transformed correlations is always less biased than when untransformed correlations are used. However, Strube (1988) noted that Silver and Dunlap had incorrectly ignored the effect of the number of studies in the analysis and so had based their findings on only a small number of studies. Strube (1988) showed that as the number of studies increased the overestimation of effect sizes based on Fisher transformed correlations was almost exactly equal in absolute terms to the underestimation of effect sizes found when untransformed  $r$ s were used. Strube's data indicated that the bias in effect size estimates based on transformed correlations was less than the bias in those based on untransformed correlations only when 3 or less studies were included in the meta-analysis (and even then only when these studies had sample sizes of 20 or less). It would be the exception that actual meta-analytic reviews would be based on such a small number of studies. As a final point, Hunter et al. (1996) have argued that when population correlations are the same for studies in the meta-analysis (the homogenous case) then results based on

transformed correlations should be within rounding error of those based on untransformed values.

The second difference is in the equations used to estimate the standard error. If we compare the random-effects model described by Hedges and Vevea (1998) to Hunter and Schmidt's, the estimates of standard error are quite different. Hedges and Vevea (1998) have suggested that Hunter and Schmidt 'advocate the use of suboptimal weights that correspond to the fixed-effects weights, presumably because they assume that  $\tau^2$  [the between-study variance] is small' (p. 493, parentheses added). Therefore, if the between-study variance is not small, the Hunter and Schmidt method will underestimate the standard error and hence overestimate the  $z$ -score associated with the mean (Hedges & Vevea, 1998). However, Hedges and Vevea's (1998) estimate of the between-study variance is truncated (because negative values lead to the assumption that  $\tau^2 = 0$ ), and so when there are only a small number of studies in the meta-analysis the estimate of between-study variance will be biased and the weights used to calculate the average effect size (and its significance) will be biased also.

Johnson et al. (1995) used a single database to compare the Hedges-Olkin (fixed-effect), Rosenthal-Rubin and Hunter-Schmidt meta-analytic methods. By manipulating the characteristics of this database Johnson et al. looked at the effects of the number of studies compared, the mean effect size of studies, the mean number of participants per study and the range of effect sizes within the database. In terms of the outcomes of each meta-analysis, they looked at the resulting mean effect size, the significance of this effect size, homogeneity of effect sizes, and prediction of effect

sizes by a moderator variable. Their results showed convergence of the methods in terms of the mean effect size and estimates of the heterogeneity of effect sizes. However, the significance of the mean effect size differed substantially across meta-analytic methods. Specifically, the Hunter and Schmidt method seemed to reach more conservative estimates of significance (and hence wider confidence intervals) than the other two methods. Johnson et al. concluded that Hunter and Schmidt's method should be used only with caution.

Johnson et al.'s study provides some of the only comparative evidence to suggest that some meta-analytic methods for combining correlations should be preferred over others (although Overton, 1998, has investigated moderator variable effects across methods); however, although their study clearly provided an excellent starting point at which to compare methods, there were some limitations. First, Schmidt and Hunter (1999) have criticised Johnson et al.'s work at a theoretical level claiming that the wrong estimate of the standard error of the mean effect size was used in their calculation of its significance. Schmidt and Hunter went on to show that when a corrected estimate was used, estimates of the significance of the mean effect size should be comparable to the Hedges and Olkin and Rosenthal and Rubin methods. Therefore, theoretically the methods should yield comparable results. Second, Johnson et al. applied Hedges and Olkin's method for  $\underline{d}$  (by first converting each correlation coefficient from  $\underline{r}$  to  $\underline{d}$ ). Hedges and Olkin (and Hedges & Vevea, 1998) provide methods for directly combining  $\underline{r}$ s (without converting to  $\underline{d}$ ) and so this procedure did not represent what researchers would actually do. Finally, the circumstances under which the three procedures were compared were limited to a

single database that was manipulated to achieve the desired changes in the independent variables of interest. This creates two concerns: (1) the conclusions drawn might be a product of the properties of the data set used (because, for example, adding or subtracting a fixed integer from each effect size allowed Johnson et al. to look at situations in which the mean effect size was higher or lower than in the original database; however, the relative strength of each effect size remained constant throughout); and (2) the data set assumed a fixed population effect size and so no comparisons were made between random-effects models. A follow-up study is needed in which Monte Carlo data simulations are used to expand Johnson et al.'s work.

### **Rationale and Predictions**

Having reviewed the procedures to be compared, some predictions can be made about their relative performance. Although there has been much theoretical debate over the efficacy of the meta-analytic methods (see Johnson et al., 1995; Schmidt & Hunter, 1999; Hedges & Vevea, 1998; Hunter and Schmidt, in press), this study aims to test the arguments empirically. The rationale is that in meta-analysis, researchers combine results from different studies to try to ascertain knowledge of the effect size in the population. Therefore, if data are sampled from a population with a known effect size, we can assess the accuracy of each method by comparing the significance of the mean effect size from the each method against the known effect in the population. In the null case (population effect size,  $\rho = 0$ ) we expect to find nonsignificant results from each meta-analysis. To be precise, with the nominal Type

I error rate set at  $\alpha = 0.05$  the expectation is that only 5% of mean effect sizes should be significant. With the population effect size set above zero the proportion of mean effect sizes that are significant represents the power of each method (assuming that the Type I error is controlled).

A number of predictions can be made based on the arguments within the literature.

1. Based on Hunter et al. (1996) and Hunter and Schmidt (1990), it is predicted that methods incorporating transformed effect size estimates should show an upward bias in their estimates of the mean effect size. This bias should be relatively small when population effect sizes are fixed (homogenous case) but larger when population effect sizes are variable (the heterogeneous case).
2. As the population value of  $\rho$  becomes further from zero, the sampling distribution of  $\rho_s$  becomes skewed and Fisher's transformation is used to normalise this sampling distribution. Therefore, theoretically Hunter and Schmidt's method should become less accurate as the effect size in the population increases (especially for small sample sizes). Conversely, techniques based on Fisher's transformation should become more accurate with larger effect sizes in the population. However, Strube (1988) and Hunter et al. (1996) have shown equivalent but opposite biases in methods based on transformed and untransformed correlation coefficients when more than a few studies are included in the meta-analysis. It is expected that the current study will replicate these later findings.
3. Contrary to Johnson et al. (1995) finding that Hunter and Schmidt's method yields conservative estimates of the significance of the mean effect size, it is

predicted that estimates of significance will be comparable across methods (because this study is based on the corrected formulae reported by Schmidt and Hunter, 1999).

4. The estimates of between-study variance in Hedges and colleagues' random-effects model are biased for small numbers of studies. As such, it is predicted that this method will be less accurate when only small numbers of studies are included in the meta-analysis.

## **STUDY 1: THE HOMOGENOUS CASE**

Two Monte Carlo studies were conducted to investigate the effect of various factors on the average effect-size, the corresponding significance value, and the homogeneity of effect-sizes tests. The first study looked at the homogenous case and the second the heterogeneous case. In both studies the general approach was the same: (1) A pseudo-population was created in which the effect size was known (homogenous case) or in which the population effect size was sampled from a normal distribution of effect sizes with a known average (heterogeneous case); (2) samples of various sizes were taken from that population and the correlation coefficient calculated and stored (these samples can be thought of as studies in a meta-analysis); (3) when a specified number of these studies had been taken different meta-analytic techniques were applied (for each technique, average effect size, the  $Z$  value and test of homogeneity was calculated); and (4) the techniques were compared to see the effect of the number of studies in the meta-analysis, and the



relative size of those studies. Each of these steps will now be discussed in more detail.

## **Method**

Both studies were run using GAUSS. In the homogenous case, a pseudo population was set up in which there was a known effect size (correlation between variables). This was achieved using the **A** matrix procedure described by Mooney (1997) in which the correlation between two randomly generated normally distributed variables is set using the Choleski decomposition of a fixed correlation matrix. Five different pseudo-populations were used in all: ones in which there were no effect ( $\rho = 0$ ), a small effect-size ( $\rho = 0.1$ ), a moderate effect size ( $\rho = 0.3$ ), a large effect size ( $\rho = 0.5$ ), and a very large effect ( $\rho = 0.8$ ). These effect sizes were based on Cohen's (1988, 1992) guidelines for a small, medium and large effect (in terms of correlation coefficients). For each Monte Carlo trial a set number of studies were taken from a given pseudo-population and average effect sizes and measures of homogeneity of effect sizes calculated. The Type I error rate or test power were estimated from the proportion of significant results over 100,000 Monte Carlo trials.

### **Number of Studies**

The first factor in the Monte Carlo study was the number of studies used in the meta- analysis. This factor varied systematically from 5 to 30 studies<sup>5</sup> in increments of 5. Therefore, for the first set of Monte Carlo trials, the program took 5 random studies from the pseudo-population on each trial. The correlation coefficients of the studies were used to calculate the mean effect size (and other statistics) using each of

the methods described. The program stored the mean effect size, and a counter was triggered if the average effect size was significant (based on the associated  $z$ -score). A second counter was triggered if the test of homogeneity was significant. Having completed this task, the next trial was run until 100,000 Monte Carlo trials were completed, after which the program saved the information, reset the counters, increased the number of studies in the meta-analysis by 5 (so the number of studies became 10) and repeated the process. The program stopped incrementing the number of studies once 30 studies was reached.

### Average Sample Size

The second factor to be varied was the average sample size of each study in the meta-analysis. This variable was manipulated to see whether the three methods differed across different sample sizes. In most real-life meta-analyses study sample sizes will not be equal and so to model reality sample sizes were drawn from a normal distribution of possible sample sizes, with the mean of this distribution being systematically varied. So, rather than fixing the sample size at a constant value (e.g. 40), sample sizes were randomly taken from a distribution with a fixed mean (in this case 40) and a standard deviation of a quarter of the mean (in this case 10). For each Monte Carlo trial, the sample size associated with the resulting  $\bar{r}$  was stored in a separate vector for use in the meta-analysis calculations.

Values of the average sample size were set using estimates of the sample size necessary to detect small, medium and large effects in the population. Based on Cohen (1988) the sample size needed to detect a small effect is 150, to detect a

medium size effect a sample of about 50 is needed and to detect a large effect a sample size of 25 will suffice. As such, the original average sample size was set at 20. Once the program has completed all computations for this sample size, the sample size was multiplied by 2 (average  $\bar{n} = 40$ ) and the program looped through all calculations again. The average sample size was then multiplied by 2 again (average  $\bar{n} = 80$ ) and so on to a maximum average sample size of 160. These sample sizes are logical because, *ceteris paribus*, the smallest average sample size (20) is big enough for only a very large effect ( $\rho = 0.8$ ) to be detected. The next sample size (40) should enable both a very large and a slightly smaller effect ( $\rho = 0.5$ ) to be detected. The next sample size (80) should be sufficient to detect all but the smallest effect sizes and the largest sample size (160) should detect all sizes of population effect sizes.

### Design

The overall design was a four factor 5 (Population effect size: 0.0, 0.1, 0.3, 0.5, 0.8)  $\times$  4 (average sample size: 20, 40, 80, 160)  $\times$  6 (Number of studies: 5, 10, 15, 20, 25, 30)  $\times$  2 (method of analysis: Hedges & Olkin/Rosenthal & Rubin vs. Hunter & Schmidt) design with the method of analysis as a repeated measure. For each level of population effect size there were 24 combinations of the average sample size and number of studies. For each of these combinations 100,000 Monte Carlo trials were used (100 times as many as the minimum recommended by Mooney, 1997) so, each cell of the design contained 100,000 cases of data. Therefore, 2,400,000 samples of data were simulated per effect size, and 12 million in the whole study.

## Results

Table 1 shows the mean effect size from the two methods when the average sample size and number of samples in the meta-analysis are varied. For the null case, all three techniques produce accurate estimates of the population effect size. As the effect size in the population increases the Hedges and Olkin/Rosenthal and Rubin method tends to slightly overestimate the population effect size whereas the Hunter and Schmidt method underestimates it. This finding was predicted because these two methods differ in their use of transformed effect size estimates. The degree of bias appears to be virtually identical when rounded to two decimal places.

<b>Insert Table 1 About Here</b>
----------------------------------

More interesting are the data presented in Table 2, which shows the proportion of significant results arising from the  $Z$  score associated with the mean effect size. This table also includes separate values for the Rosenthal-Rubin method (because it differs from the Hedges-Olkin method in terms of how significance is established). In the null case, these proportions represent the Type I error rate for the three methods. Using a nominal  $\alpha$  of 0.05 it is clear that the Hedges-Olkin method keeps tight control over the Type I error rate (this finding supports data presented by Hedges & Vevea, 1998). The Hunter-Schmidt method does not control the Type I error rate in the homogenous case, although for a large total sample size (i.e. as the number of studies in the meta-analysis and the average sample size of each study increases) the Type I error rate is better controlled ( $\alpha \approx 0.06$ ). However, for small numbers of studies and small average sample sizes the Type I error rate is around thrice the desirable level. The Rosenthal-Rubin method keeps fairly tight control of the Type I error rate with

error rates falling between those observed using the Hedges-Olkin and the Hunter-Schmidt methods. For the remainder of Table 2 (in which an effect exists within the population), the proportions represent the power of the tests assuming that the Type I error rate is controlled (as such the values for the Hedges-Olkin and Rosenthal-Rubin methods are estimates of the power of the test). A proportion of 0.8 generally represents a high level of power in the social sciences (Cohen, 1988, 1992). The power of the meta-analysis will increase as the total sample size increases and so as both the number of studies, and their respective sample sizes increase, we expect a concomitant increase in power. The methods advocated by Hedges and Olkin, and Rosenthal and Rubin both yield high levels of power (greater than 0.8) except when the population effect size is small ( $\rho = 0.1$ ) and the total sample size is relatively small. For example, when number of studies in the meta-analysis is small (5 studies) a high level of power is achieved only when the average study sample size is 160, similarly, regardless of the number of studies, a high level of power is not achieved when the average study sample size is only 20, and when the average sample size is 40, a high degree of power is achieved only when there are more than 20 studies. For all other population effect sizes ( $\rho > 0.1$ ) the probability of detecting a genuine effect is greater than 0.8. For the Hunter-Schmidt method power estimations cannot be made because the Type I error rate is not controlled, nevertheless, the values in Table 2 are comparable to those for the other two methods.

<p><b>Insert Table 2 About Here</b></p>
-----------------------------------------

Table 3 shows the proportion of significant tests of homogeneity of effect sizes. In this study, the population effect sizes were fixed (hence homogenous); therefore,

these tests of homogeneity should yield nonsignificant results. The proportions in Table 3 should, therefore, be close to the nominal  $\alpha$  of 0.05. For small to medium effect sizes ( $\rho \leq 0.3$ ), both methods control the Type I error rate under virtually all conditions. For larger population effect sizes ( $\rho \geq 0.5$ ) the Hedges-Olkin/Rosenthal-Rubin method controls the Type I error rate to within rounding error of the nominal  $\alpha$ . However, the Hunter-Schmidt method begins to deviate substantially from the nominal  $\alpha$  when the average sample size is small ( $< 40$ ) and this deviation increases as the number of studies within the meta-analysis increases. These results conform to accepted statistical theory (see prediction 2) in that the benefit of transformed effect sizes is increasingly apparent as the population effect size increases.

<p><b>Insert Table 3 About Here</b></p>
-----------------------------------------

### Summary

To sum up, study 1 empirically demonstrated several things. (1) Both meta-analytic methods yield comparable estimates of population effect sizes; (2) The Type I error rates were well controlled for the Hedges-Olkin and Rosenthal-Rubin methods in all circumstances, however, the Hunter-Schmidt method seemed to produce liberal significance tests that inflated the observed error rate above the nominal  $\alpha$ ; (3) the Hedges-Olkin and Rosenthal-Rubin methods yielded power levels above 0.8 for medium and large effect sizes, but not for small effect sizes when the number of studies, or average sample size were relatively small; (4) Type I error rates for tests of homogeneity of effect sizes were equally well controlled by the two methods when

population effect sizes were small to medium but better controlled by the Hedges-Olkin/Rosenthal-Rubin method when effect sizes were large.

## STUDY 2: THE HETEROGENEOUS CASE

### Method

The method for the heterogeneous case was virtually identical to that of the homogenous case: both the number of studies and the average sample size were varied in the same systematic way. However, in this study, population effect sizes were not fixed. A normal distribution of possible effect sizes was created (a superpopulation) from which the population effect size for each study in a meta-analysis was sampled. As such, studies in a meta-analysis came from populations with different effect sizes. To look at a variety of situations, the mean effect size of the superpopulation ( $\bar{\rho}$ ) was varied to be 0 (the null case), 0.1, 0.3, 0.5, and 0.8. The standard deviation of the superpopulation was set at 0.16 because (a) for a medium population effect size ( $\rho = 0.3$ ) this represents a situation in which 95% of population effect sizes will lie between 0 (no effect) and 0.6 (strong effect), and (b) Barrick and Mount (1991) found this to be the standard deviation of population correlations in a large meta-analysis and so it represents a realistic estimate of the standard deviation of population correlations of real-world data (see Hunter & Schmidt, in press). The methods used to combine correlation coefficients in this study were the Hunter-Schmidt method and Hedges and colleagues' random effects model. As in study 1,

100,000 Monte Carlo trials were used for each combination of average sample size, number of studies in the meta-analysis, and average population effect size.

## Results

Table 4 shows the mean effect size from the two methods when the average sample size and number of studies in the meta-analysis are varied. In the null case both methods produce accurate estimations of the population effect size. When there is an effect in the population the Hedges and colleagues' method uniformly overestimates and the Hunter-Schmidt method uniformly underestimates the population effect size. This is expected from prediction 1. The overestimation from the Hedges and colleagues method is substantial and is typically around 0.1–0.2 greater than the actual average population effect size for medium to large effects. In contrast the underestimation of the Hunter-Schmidt method is relatively small (typically within rounding error of the actual average population value). However, as the average population effect size increases, so does the margin of error in the Hunter-Schmidt estimations and at very large average population effect sizes ( $\bar{\rho} = 0.8$ ) the magnitude of the underestimation of this method is equivalent to the overestimation of Hedges and colleagues' method. This finding was predicted because at larger effect sizes the benefit of the  $r$  to  $z$  transformation should be more apparent (prediction 2), although even at very high effect sizes the bias from transforming  $r$  to  $z$  is the same but opposite to that of not transforming  $r$ .

<b>Insert Table 4 About Here</b>
----------------------------------



Table 5 shows the proportion of significant results arising from the  $z$ -score associated with the mean effect sizes in Table 4. In the null case, these proportions represent the Type I error rate for the two methods. Using a nominal  $\alpha$  of 0.05, it is clear that both techniques lead to inflated Type I error rates when there are 15 or less studies in the meta-analysis (although the Hedges and colleagues' method retains better control than the Hunter-Schmidt method). Control of the Type I error rate improves in both methods as the total sample size increases, and when the meta-analysis includes a large number of studies (30), Hedges and colleagues' method produces error rates within rounding distance of the nominal  $\alpha$ -level. Even for large numbers of studies, the Hunter-Schmidt method inflates the Type 1 error rate.

For the remainder of the table (in which an effect exists within the population), the proportions displayed represent the power of the tests assuming that the Type I error rate is controlled. Given that neither method has absolute control over the Type I error rate these values need to be interpreted cautiously. What is clear is that the two methods yield very similar results: for a small average population effect size ( $\bar{\rho} = 0.1$ ) the probability of detecting an effect is under 0.3 for both methods. High probabilities of detecting an effect ( $> 0.8$ ) are achieved only for large average population effect sizes ( $\bar{\rho} \geq 0.5$ ) or for medium effect sizes ( $\bar{\rho} = 0.3$ ) when there are a relatively large number of samples (20 or more). The only substantive discrepancy between the methods is that the values for Hedges and colleagues' method are lower when there are only 5 studies in the meta-analysis, which was predicted (prediction

4). This difference is negligible when the average population effect size is very large ( $\bar{\rho} = 0.8$ ).

**Insert Table 5 About Here**

Table 6 shows the proportion of significant tests of homogeneity of effect sizes. In this study, the population effect sizes were heterogeneous; therefore, these tests should yield significant results. The proportions in Table 6, therefore, represent the power of the tests to detect variability in effect sizes assuming that the Type I error rate is controlled. This study does not present data to confirm that the methods control the Type I error rate (which would require that these tests be applied to the homogenous case); nevertheless, for all average population effect sizes the two methods yield probabilities of detecting an effect greater than 0.8 with samples of 40 or more regardless of the number of studies in the meta-analysis. Even at small sample sizes and numbers of studies, the proportion of tests that correctly detected genuine variance between population parameters is close to 0.8 (the lowest probability being 0.704) and comparable between methods.

**Insert Table 6 About Here**

### Summary

To sum up, study 2 empirically demonstrated several interesting findings. (1) The Hunter-Schmidt method produces the most accurate estimates of population effect sizes when population effect sizes are variable but the benefit of this method is lost when the average population effect size is very large ( $\bar{\rho} = 0.8$ ); (2) The Type I error

rates were not controlled by either method when 15 or less studies were included in the meta-analysis (although the Hedges-Olkin method was better in this respect), however, as the total sample size increased the Type I error rate was better controlled for both methods; (3) although flawed by the lack of control of the Type I error rate, the potential power of both techniques was less than 0.3 when the average population effect size was small; (4) for large average population effect sizes the three techniques were comparable for probable test power, but for small numbers of studies in the meta-analysis, Hedges and colleagues' method yielded lower power estimates; (5) power rates for tests of homogeneity of effect sizes were comparable for both techniques in all circumstances.

## **CONCLUSIONS**

This study presents the results of a thorough simulation of conditions that might influence the efficacy of different methods of meta-analysis. In an attempt to develop Johnson et al.'s (1995) work, this study used Monte Carlo simulation rather than manipulation of a single data set. In doing so, these data provide a broader insight into the behaviour of different meta-analytic procedures in an applied context (rather than the theoretical context of Hunter & Schmidt, in press; Schmidt & Hunter, 1999). Several predictions were supported. (1) Prediction 1 was substantiated in that the Hedges-Olkin and Rosenthal-Rubin methods (using transformed effect size estimates) led to upward biases in effect size estimates. These biases were negligible in the homogenous case but substantial in the heterogeneous case. (2) Prediction 2

was also substantiated with the Hunter-Schmidt method underestimating population effect sizes. This bias increased as the population effect sizes increased (as predicted). However, this bias was negligible in the homogeneous case and was less severe than the Hedges-Olkin method in the Heterogeneous case. (3) Results for prediction 3 were complex. The Hedges-Olkin method best controlled the Type I error rate and the Hunter-Schmidt method led to the greatest deviations from the nominal  $\alpha$ . However, unlike Johnson et al.'s who found that the Hunter-Schmidt method was too conservative, this study showed that the Hunter and Schmidt method (using their revised formula) was too liberal—too many null results were significant). These results also contradict the theoretical observations of Hunter and Schmidt (in press) and Schmidt and Hunter (1999). (4) Prediction 4 was also supported in that Hedges and colleagues' method was slightly biased when only 5 studies were included in the meta-analysis in the heterogeneous case.

In summary, in the homogenous case Hedges-Olkin and Rosenthal-Rubin methods perform best in terms of significance tests of the average effect size: contrary to Johnson et al. (1995), the present results indicate that the Hunter-Schmidt method is too liberal in the homogenous case (not too conservative) but this means that the method should, nevertheless, be applied with caution in these circumstances. In terms of estimates of effect size and homogeneity of effect size tests there are few differences between Hedges-Olkin/Rosenthal-Rubin methods and that of Hunter and Schmidt. In the heterogeneous case, the Hunter-Schmidt method yields the most accurate estimates of population effect size across a variety of situations. The most surprising result was that neither the Hunter-Schmidt nor Hedges and colleagues'

method controlled the Type I error rate in the heterogeneous case when 15 or fewer studies were included in the meta-analysis. As such, in the heterogeneous case researchers cannot be confident about the tests they use unless the number of studies being combined (and hence the total sample size) is very large (at least 30 studies in the meta-analysis for Hedges and colleagues' method and more for the Hunter-Schmidt method). In addition, the probabilities of detecting a small effect in the heterogeneous case were very small, and for medium effect sizes were small when 10 or less studies were in the meta-analysis. Given that the heterogeneous case is more representative of real-world data (National Research Council, 1992, Osburn and Callender, 1992) the implication is that meta-analytic methods for combining correlation coefficients may be relatively insensitive to detecting small effects in the population. As such, genuine effects may be overlooked. However, this conclusion must be qualified: when 15 or less studies are included in the meta-analysis neither random effects model controls the Type I error rate, as such accurate power levels cannot be estimated. As such, the finding that the probabilities of detecting medium population effect sizes ( $\underline{\bar{\rho}} = 0.3$ ) are low for less than 15 studies is, at best, tentative. Nevertheless, for small population effect sizes ( $\underline{\bar{\rho}} = 0.1$ ), even when Type I error rates are controlled (the Hedges and Colleagues' method when 20 or more studies are included in the meta-analysis) the power of the random-effects model is relatively small (average power across all factors is 0.209).

### Using Meta-Analysis for Correlations

There are many considerations when applying techniques to combine correlation coefficients. The first is whether the researcher wishes to make conditional or unconditional inferences from the meta-analysis, or in other terms, whether the researcher wishes to assume that the population effect size is fixed or variable. As already mentioned, it is more often the case that population effect sizes are variable (National Research Council, 1992, Osburn and Callender, 1992) and that the assumption of fixed population effect sizes is tenable only if a researcher does not wish to generalise beyond the set of studies within a meta-analysis (Hedges & Vevea, 1998; Hunter & Schmidt, in press). One practical way to assess whether population effect sizes are likely to be fixed or variable is to use the tests of homogeneity of study effect sizes associated with the three methods of meta-analysis. If this test is non-significant then it can be argued that population effect sizes are also likely to be homogenous (and hence fixed to some extent). However, these tests typically have low power to detect genuine variation in population effect sizes (Hedges & Olkin, 1985; National Research Council, 1992) and so they can lead researchers to conclude erroneously that population effect sizes are fixed. The present data suggest that the test of homogeneity of effect sizes advocated by Hedges-Olkin/Rosenthal-Rubin and the method suggested by Hunter and Schmidt have relatively good control of Type I errors when effect sizes are, in reality, fixed. When effect sizes are, in reality, variable both Hedges and colleagues' method and the Hunter-Schmidt method produce equivalent estimates of power (although when the average effect size is large and the average sample size is less than 40 the Hunter-Schmidt method loses control of the

Type I error rate). However, in this later case these detection rates are difficult to interpret because there are no simulations in the current study to test whether the random-effects homogeneity tests control the Type I error rate in the fixed case (when population effect sizes are, in reality, the same across studies).

The second issue is whether the researcher wishes to accurately estimate the population effect size, or accurately test its significance. In the homogenous case, all method yield very similar estimates of the population effect size. However, in the heterogeneous case the Hunter-Schmidt method produces more accurate estimates except when the average population effect size is very large ( $\bar{\rho} = 0.8$ ). In terms of testing the significance of this estimate, Hedges and colleagues' method keeps the best control over Type I errors in both the homogenous and heterogeneous case, however, in the heterogeneous case neither this method or the Hunter-Schmidt method actually controls the Type I error rate acceptably when less than 15 studies are included in the meta-analysis.

Third, the researcher has to consider controlling for other sources of error. It is worth remembering that small statistical differences in average effect size estimates and the like may be relatively unimportant compared to other forms of bias such as unreliability of measures. Hunter & Schmidt (1990) discuss ways in which these biases can be accounted for and the experienced meta-analyst should consider these issues when deciding upon a technique. The Hunter-Schmidt method used in the present paper is only the simplest form of this method and so does not reflect the full method adequately. Despite its relative shortcomings in the homogenous case, the addition of procedures for controlling other sources of bias may make this method

very attractive in situations in which the researcher can estimate and control for these other confounds. However, further research is needed to test the accuracy of the adjustments for error sources proposed by Hunter and Schmidt.

### Final Remarks

This study has shown that the Hunter-Schmidt method tends to provide the most accurate estimates of the mean population effect size when effect sizes are heterogeneous, which is the most common case in meta-analytic practice. In the heterogeneous case, Hedges and colleagues' method tended to overestimate effect sizes by about 15-45%, whereas the Hunter-Schmidt method tended to underestimate it by a smaller amount (about 5-10%), and then only when the population average correlation exceeded 0.5. In terms of the Type I error rate for the significance tests associated with these estimates Hedges and colleagues' method does control this error rate in the homogenous case. The most surprising finding is that neither random-effects method controls the Type I error rate in the heterogeneous case (except when a large number of studies are included in the meta-analysis) — although Hedges and colleagues' method inflates the Type I error rate less than the Hunter-Schmidt method. Given that the National Research Council (1992) and others have suggested that the heterogeneous case is the rule rather than the exception, this implies that estimates and significance tests from meta-analytic studies containing less than 30 samples should be interpreted very cautiously. Even then, random-effects methods seem poor at detecting small population effect sizes. Further work should examine the efficacy of other random-effect models of meta-analysis such as



Multilevel Modelling (Goldstein, 1995, Goldstein, Yang, Omar, Turner & Thompson, 2000).

## FOOTNOTES

---

<sup>1</sup> The Web of Science (WoS) was used (<http://wos.mimas.ac.uk>).

<sup>2</sup> In reality it is possible to combine fixed and random effects conceptualizations to produce a mixed model. For the purpose of this study the mixed model is ignored but the interested reader is referred to Hedges (1992).

<sup>3</sup> Although Hunter, Schmidt & Jackson (1982) originally developed this method, Hunter and Schmidt (1990) provide an updated and more comprehensive exposition of the technique.

<sup>4</sup> In fact the equation for the mean effect size (see equation 11) implies a fixed-effects model because the use of  $\underline{n}_i$  as a weight assumes homogeneity (and indeed Hunter and Schmidt, 1990, p. 100 assert the homogeneity assumption). However, in more recent work (Schmidt & Hunter, 1999; Hunter & Schmidt, in press) the authors have been quite explicit in labelling their model as random-effect.

<sup>5</sup> The number of studies in real meta-analytic studies is likely to exceed 30 and would rarely be as small as 5, nevertheless these values are fairly typical of moderator analysis in meta-analysis.

## TABLES

- Table 1: Table to show the mean effect size,  $\bar{r}$ , for the two methods of meta-analysis for different average sample sizes, different numbers of studies in the meta-analysis, and different levels of population effect size (homogenous case).
- Table 2: Table to show the proportion of significant tests of the mean effect size for different numbers of samples, different average sample sizes, and different levels of population effect size (homogenous case).
- Table 3: Table to show the proportion of significant tests of homogeneity of sample effect sizes for different numbers of samples, different average sample sizes, and different levels of population effect size (homogenous case).
- Table 4: Table to show the mean effect size,  $\bar{r}$ , for the two methods of meta-analysis for different average sample sizes, different numbers of studies in the meta-analysis, and different levels of population effect size (heterogeneous case).
- Table 5: Table to show the proportion of significant tests of the mean effect size for different numbers of samples, different average sample sizes, and different levels of population effect size (heterogeneous case).
- Table 6: Table to show the proportion of significant tests of homogeneity of sample effect sizes for different numbers of samples, different average sample sizes, and different levels of population effect size (heterogeneous case).

		Hedges-Olkin/Rosenthal-Rubin				Hunter & Schmidt			
		Average Sample Size							
		20	40	80	160	20	40	80	160
		$\rho = 0.0$							
Number of Studies in the Meta-Analysis	5	-0.0002	0.0001	-0.0001	0.0001	-0.0001	0.0001	-0.0001	0.0001
	10	-0.0002	0.0000	0.0000	-0.0001	-0.0002	0.0000	0.0000	-0.0001
	15	0.0001	-0.0001	0.0000	0.0000	0.0001	-0.0001	0.0000	0.0000
	20	0.0000	0.0000	-0.0001	0.0000	0.0000	0.0000	-0.0001	0.0000
	25	-0.0001	-0.0001	0.0000	0.0000	-0.0001	-0.0001	0.0000	0.0000
	30	0.0000	-0.0001	0.0000	0.0000	0.0000	-0.0001	0.0000	0.0000
		$\rho = 0.1$							
	5	0.1016	0.1007	0.1004	0.1002	0.0974	0.0986	0.0994	0.0997
	10	0.1024	0.1010	0.1003	0.1002	0.0976	0.0987	0.0992	0.0996
	15	0.1024	0.1009	0.1006	0.1003	0.0974	0.0985	0.0994	0.0997
	20	0.1026	0.1011	0.1006	0.1002	0.0975	0.0986	0.0994	0.0996
	25	0.1027	0.1013	0.1005	0.1003	0.0975	0.0988	0.0993	0.0997
	30	0.1026	0.1012	0.1006	0.1003	0.0974	0.0987	0.0994	0.0997
		$\rho = 0.3$							
	5	0.3047	0.3022	0.3009	0.3003	0.2930	0.2965	0.2981	0.2990
	10	0.3057	0.3028	0.3014	0.3007	0.2925	0.2964	0.2983	0.2992
	15	0.3065	0.3032	0.3014	0.3008	0.2927	0.2966	0.2982	0.2991
	20	0.3065	0.3032	0.3015	0.3008	0.2925	0.2965	0.2982	0.2992
	25	0.3070	0.3033	0.3016	0.3008	0.2929	0.2965	0.2983	0.2991
	30	0.3071	0.3033	0.3016	0.3008	0.2928	0.2964	0.2982	0.2991
		$\rho = 0.5$							
	5	0.5066	0.5031	0.5013	0.5007	0.4903	0.4953	0.4975	0.4988
	10	0.5085	0.5040	0.5020	0.5010	0.4902	0.4952	0.4977	0.4988
	15	0.5090	0.5044	0.5020	0.5010	0.4899	0.4953	0.4976	0.4988
	20	0.5094	0.5046	0.5021	0.5011	0.4899	0.4953	0.4975	0.4989
	25	0.5098	0.5045	0.5023	0.5011	0.4901	0.4951	0.4977	0.4988
	30	0.5099	0.5046	0.5022	0.5011	0.4901	0.4952	0.4975	0.4988
		$\rho = 0.8$							
	5	0.8048	0.8021	0.8011	0.8005	0.7919	0.7960	0.7981	0.7991
	10	0.8063	0.8030	0.8015	0.8007	0.7918	0.7962	0.7981	0.7991
	15	0.8070	0.8033	0.8017	0.8008	0.7920	0.7962	0.7982	0.7991
	20	0.8071	0.8033	0.8017	0.8008	0.7918	0.7961	0.7982	0.7991
	25	0.8073	0.8034	0.8017	0.8008	0.7919	0.7961	0.7981	0.7991
	30	0.8075	0.8036	0.8017	0.8008	0.7919	0.7962	0.7982	0.7991

		Hedges & Olkin				Rosenthal & Rubin				Hunter & Schmidt			
		Average Sample Size											
		20	40	80	160	20	40	80	160	20	40	80	160
		$\rho = 0.0$											
Number of Studies in the Meta-Analysis	5	0.051	0.051	0.049	0.050	0.056	0.054	0.052	0.050	0.156	0.156	0.153	0.155
	10	0.051	0.051	0.050	0.050	0.054	0.053	0.051	0.050	0.098	0.098	0.097	0.096
	15	0.050	0.051	0.050	0.049	0.055	0.052	0.052	0.051	0.080	0.081	0.079	0.080
	20	0.050	0.051	0.049	0.051	0.056	0.050	0.052	0.050	0.071	0.073	0.071	0.073
	25	0.050	0.049	0.049	0.050	0.056	0.054	0.051	0.050	0.067	0.067	0.066	0.067
	30	0.051	0.049	0.051	0.050	0.056	0.052	0.050	0.050	0.064	0.063	0.065	0.063
		$\rho = 0.1$											
	5	0.155	0.282	0.508	0.805	0.167	0.288	0.513	0.810	0.283	0.417	0.618	0.845
	10	0.270	0.496	0.799	0.978	0.284	0.505	0.804	0.979	0.343	0.557	0.819	0.977
	15	0.378	0.666	0.930	0.998	0.395	0.675	0.931	0.997	0.424	0.692	0.931	0.998
	20	0.477	0.789	0.978	1.000	0.496	0.798	0.976	1.000	0.510	0.800	0.976	1.000
	25	0.568	0.871	0.993	1.000	0.585	0.873	0.994	1.000	0.593	0.873	0.992	1.000
	30	0.646	0.923	0.998	1.000	0.665	0.922	0.998	1.000	0.661	0.924	0.998	1.000
		$\rho = 0.3$											
	5	0.836	0.989	1.000	1.000	0.847	0.989	1.000	1.000	0.868	0.986	1.000	1.000
	10	0.985	1.000	1.000	1.000	0.988	1.000	1.000	1.000	0.983	1.000	1.000	1.000
	15	0.999	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.999	1.000	1.000	1.000
	20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\rho = 0.5$											
	5	0.999	1.000	1.000	1.000	0.999	1.000	1.000	1.000	0.998	1.000	1.000	1.000
	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\rho = 0.8$											
	5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

		Hedges-Olkin/Rosenthal-Rubin				Hunter & Schmidt			
		Average Sample Size							
		20	40	80	160	20	40	80	160
		$\rho = 0.0$							
Number of Studies in the Meta-Analysis	5	0.053	0.050	0.052	0.050	0.047	0.047	0.051	0.050
	10	0.053	0.051	0.052	0.051	0.042	0.046	0.049	0.050
	15	0.055	0.052	0.051	0.051	0.041	0.045	0.047	0.049
	20	0.054	0.052	0.051	0.050	0.040	0.046	0.048	0.048
	25	0.054	0.052	0.051	0.051	0.039	0.045	0.048	0.050
	30	0.055	0.053	0.052	0.051	0.040	0.045	0.048	0.049
		$\rho = 0.1$							
	5	0.053	0.052	0.050	0.051	0.048	0.049	0.049	0.050
	10	0.053	0.051	0.052	0.050	0.042	0.046	0.049	0.049
	15	0.053	0.051	0.052	0.050	0.042	0.045	0.049	0.049
	20	0.055	0.051	0.052	0.050	0.042	0.046	0.049	0.048
	25	0.053	0.052	0.052	0.050	0.040	0.046	0.048	0.048
	30	0.055	0.053	0.051	0.050	0.041	0.046	0.048	0.048
		$\rho = 0.3$							
	5	0.052	0.051	0.051	0.049	0.049	0.050	0.050	0.049
	10	0.054	0.052	0.051	0.050	0.049	0.050	0.051	0.050
	15	0.052	0.052	0.051	0.050	0.049	0.051	0.051	0.050
	20	0.054	0.051	0.051	0.051	0.049	0.051	0.051	0.050
	25	0.052	0.052	0.051	0.051	0.050	0.051	0.051	0.051
	30	0.052	0.052	0.051	0.050	0.050	0.051	0.051	0.050
		$\rho = 0.5$							
	5	0.050	0.050	0.050	0.050	0.054	0.051	0.051	0.050
	10	0.051	0.051	0.050	0.051	0.061	0.055	0.053	0.053
	15	0.051	0.051	0.050	0.050	0.067	0.058	0.055	0.052
	20	0.051	0.051	0.051	0.051	0.069	0.061	0.056	0.053
	25	0.052	0.052	0.051	0.050	0.073	0.063	0.057	0.052
	30	0.050	0.050	0.050	0.050	0.075	0.063	0.056	0.053
		$\rho = 0.8$							
	5	0.048	0.049	0.051	0.050	0.063	0.057	0.054	0.052
	10	0.048	0.050	0.050	0.051	0.090	0.071	0.060	0.056
	15	0.047	0.050	0.050	0.050	0.107	0.078	0.066	0.057
	20	0.048	0.049	0.050	0.050	0.122	0.085	0.068	0.059
	25	0.047	0.050	0.049	0.051	0.134	0.092	0.071	0.060
	30	0.046	0.049	0.050	0.049	0.144	0.095	0.073	0.060

		Hedges and colleagues				Hunter & Schmidt			
		Average Sample Size							
		20	40	80	160	20	40	80	160
		$\bar{\rho} = 0.0$							
Number of Studies in the Meta-Analysis	5	-0.001	0.000	0.001	0.000	-0.001	0.001	0.001	0.000
	10	-0.001	0.001	0.000	0.000	-0.001	0.001	0.000	0.000
	15	0.000	0.000	-0.001	0.000	0.000	0.000	0.000	0.000
	20	0.000	0.000	-0.001	0.000	0.000	0.000	0.000	0.000
	25	0.000	0.000	-0.001	0.001	0.000	0.000	0.000	0.000
	30	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		$\bar{\rho} = 0.1$							
	5	0.134	0.134	0.131	0.131	0.098	0.099	0.098	0.099
	10	0.141	0.140	0.139	0.139	0.098	0.098	0.099	0.099
	15	0.142	0.143	0.141	0.142	0.097	0.099	0.098	0.099
	20	0.143	0.144	0.142	0.143	0.097	0.099	0.098	0.099
	25	0.144	0.144	0.143	0.143	0.097	0.099	0.098	0.099
	30	0.146	0.144	0.143	0.144	0.097	0.098	0.098	0.099
		$\bar{\rho} = 0.3$							
	5	0.397	0.395	0.395	0.394	0.290	0.292	0.293	0.293
	10	0.421	0.417	0.417	0.416	0.290	0.291	0.292	0.293
	15	0.430	0.427	0.425	0.424	0.290	0.292	0.293	0.293
	20	0.434	0.431	0.429	0.429	0.290	0.292	0.292	0.293
	25	0.436	0.433	0.431	0.432	0.290	0.291	0.292	0.293
	30	0.438	0.436	0.433	0.433	0.289	0.292	0.293	0.293
		$\bar{\rho} = 0.5$							
	5	0.643	0.641	0.640	0.639	0.474	0.477	0.478	0.479
	10	0.680	0.677	0.676	0.675	0.475	0.477	0.478	0.479
	15	0.693	0.690	0.688	0.688	0.474	0.478	0.478	0.480
	20	0.699	0.696	0.694	0.695	0.474	0.477	0.478	0.479
	25	0.704	0.701	0.700	0.699	0.474	0.477	0.478	0.479
	30	0.707	0.703	0.702	0.701	0.475	0.477	0.478	0.479
		$\bar{\rho} = 0.8$							
	5	0.896	0.895	0.894	0.894	0.715	0.719	0.719	0.720
	10	0.922	0.920	0.919	0.920	0.715	0.718	0.719	0.720
	15	0.930	0.929	0.928	0.928	0.716	0.719	0.720	0.720
	20	0.934	0.933	0.932	0.932	0.716	0.718	0.720	0.720
	25	0.936	0.935	0.934	0.934	0.716	0.718	0.719	0.720
	30	0.938	0.937	0.936	0.936	0.716	0.719	0.719	0.720

		Hedges and colleagues				Hunter & Schmidt			
		Average Sample Size							
		20	40	80	160	20	40	80	160
		$\bar{\rho} = 0.0$							
Number of Studies in the Meta-Analysis	5	0.100	0.106	0.106	0.106	0.158	0.157	0.156	0.156
	10	0.069	0.069	0.070	0.069	0.097	0.098	0.098	0.096
	15	0.061	0.059	0.058	0.059	0.082	0.081	0.080	0.081
	20	0.055	0.054	0.053	0.053	0.073	0.072	0.072	0.073
	25	0.055	0.054	0.053	0.051	0.069	0.069	0.069	0.067
	30	0.051	0.051	0.050	0.049	0.065	0.065	0.065	0.064
		$\bar{\rho} = 0.1$							
	5	0.105	0.117	0.117	0.118	0.163	0.172	0.173	0.174
	10	0.116	0.121	0.124	0.124	0.159	0.167	0.172	0.174
	15	0.133	0.143	0.143	0.149	0.172	0.189	0.191	0.198
	20	0.159	0.172	0.171	0.179	0.198	0.217	0.221	0.230
	25	0.187	0.202	0.207	0.212	0.222	0.245	0.256	0.263
	30	0.219	0.232	0.237	0.244	0.253	0.275	0.284	0.295
		$\bar{\rho} = 0.3$							
	5	0.298	0.319	0.333	0.339	0.452	0.481	0.501	0.508
	10	0.478	0.505	0.525	0.534	0.608	0.651	0.676	0.690
	15	0.665	0.697	0.716	0.724	0.754	0.795	0.819	0.827
	20	0.799	0.830	0.845	0.853	0.851	0.887	0.904	0.913
	25	0.885	0.907	0.919	0.923	0.914	0.941	0.954	0.958
	30	0.933	0.949	0.956	0.959	0.952	0.970	0.978	0.981
		$\bar{\rho} = 0.5$							
	5	0.534	0.559	0.569	0.579	0.766	0.804	0.823	0.834
	10	0.864	0.883	0.893	0.900	0.941	0.960	0.969	0.972
	15	0.975	0.983	0.986	0.987	0.988	0.995	0.996	0.997
	20	0.996	0.997	0.998	0.998	0.998	0.999	1.000	1.000
	25	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000
	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\bar{\rho} = 0.8$							
	5	0.867	0.879	0.884	0.886	0.976	0.985	0.990	0.991
	10	0.999	0.999	0.999	1.000	1.000	1.000	1.000	1.000
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

		Hedges and colleagues				Hunter & Schmidt			
		Average Sample Size							
		20	40	80	160	20	40	80	160
		$\bar{\rho} = 0.0$							
Number of Studies in the Meta-Analysis	5	0.720	0.882	0.958	0.988	0.704	0.879	0.958	0.988
	10	0.925	0.989	0.999	1.000	0.913	0.988	0.999	1.000
	15	0.980	0.999	1.000	1.000	0.975	0.999	1.000	1.000
	20	0.995	1.000	1.000	1.000	0.993	1.000	1.000	1.000
	25	0.999	1.000	1.000	1.000	0.998	1.000	1.000	1.000
	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\bar{\rho} = 0.1$							
	5	0.729	0.884	0.959	0.988	0.712	0.880	0.959	0.988
	10	0.928	0.989	0.999	1.000	0.916	0.988	0.999	1.000
	15	0.982	0.999	1.000	1.000	0.977	0.999	1.000	1.000
	20	0.996	1.000	1.000	1.000	0.994	1.000	1.000	1.000
	25	0.999	1.000	1.000	1.000	0.998	1.000	1.000	1.000
	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\bar{\rho} = 0.3$							
	5	0.780	0.910	0.969	0.990	0.762	0.907	0.969	0.990
	10	0.954	0.994	0.999	1.000	0.944	0.993	0.999	1.000
	15	0.990	1.000	1.000	1.000	0.987	1.000	1.000	1.000
	20	0.998	1.000	1.000	1.000	0.997	1.000	1.000	1.000
	25	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000
	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\bar{\rho} = 0.5$							
	5	0.863	0.945	0.982	0.995	0.842	0.942	0.982	0.995
	10	0.983	0.998	1.000	1.000	0.976	0.998	1.000	1.000
	15	0.998	1.000	1.000	1.000	0.997	1.000	1.000	1.000
	20	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000
	25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		$\bar{\rho} = 0.8$							
	5	0.967	0.986	0.994	0.996	0.953	0.985	0.994	0.996
	10	0.999	1.000	1.000	1.000	0.998	1.000	1.000	1.000
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000



# REFERENCES

- Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. Personnel Psychology, 44, 1-26.
- Becker, B. J. (1996). The generalizability of empirical research results. In C. P. Benbow and D. Lubinski (Eds.), Intellectual talent: Psychological and social issues, Baltimore: John Hopkins University Press, 363-383.
- Cohen, J. (1988). Statistical power analysis for the behavioural sciences (2<sup>nd</sup> Ed.). New York: Academic Press.
- Cohen, J. (1992). A power primer. Psychological Bulletin, 112 (1), 155–159.
- Field, A. P. (1999). A bluffer's guide to meta-analysis I: Correlations. Newsletter of the Mathematical, Statistical and computing section of the British Psychological Society, 7 (1), 16-25.
- Field, A. P. (2000). Discovering statistics using SPSS for Windows: advanced techniques for the beginner. London: Sage.
- Fisher, R. A. (1928). Statistical methods for research workers (2<sup>nd</sup> Ed.). London: Oliver & Boyd.
- GAUSS for Windows 3.2.35 [Computer software]. (1998). Maple Valley, WA: Aptech Systems, Inc.
- Glass, G. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5I, 3-8.
- Glass, G., McGaw, B, & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.
- Goldstein, H. (1995). Multilevel statistical models. London: Edward Arnold.
- Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. Journal of the Royal Statistical Society Series C – Applied Statistics, 49 (3), 399-412.
- Hedges, L. V. (1992). Meta-Analysis. Journal of Educational Statistics, 17 (4), 279-296.

- Hedges, L. V. (1994a). Statistical considerations. In H. Cooper and L. V. Hedges (Eds.), Handbook of research synthesis. New York: Russell Sage Foundation, 29-38.
- Hedges, L. V. (1994b). Fixed effects models. In H. Cooper and L. V. Hedges (Eds.), Handbook of research synthesis. New York: Russell Sage Foundation, 285-300.
- Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.
- 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. (1990). Methods of Meta-analysis: correcting error and bias in research findings. Newbury Park, CA: Sage.
- Hunter, J. E., & Schmidt, F. L. (in press). Fixed effects vs. random effects meta-analysis models: implications for cumulative knowledge in Psychology. International Journal of Selection and Assessment.
- Hunter, J. E., Schmidt, F. L., & Coggin, T. D. (1996). Meta-analysis of correlations: bias in the correlation coefficient and the Fisher  $z$  transformation. Unpublished Manuscript.
- Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: cumulating research findings across studies. Beverly Hills, CA: Sage.
- Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. Journal of Applied Psychology, 80 (1), 94-106.
- Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994). Nonlinearity of range corrections in meta-analysis: Test of an improved procedure. Journal of Applied Psychology, 79 (3), 425-438.
- Mooney, C. Z. (1997). Monte Carlo Simulation (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-116). Thousand Oaks, CA: Sage.
- National Research Council (1992). Combining information: Statistical issues and opportunities for research. Washington, D.C.: National Academy Press.

- Osburn, H. G., & Callender, J. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. Journal of Applied Psychology, 77, 115-122.
- Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. Psychological Methods, 3 (3), 354-379.
- Rosenthal, R. (1991). Meta-analytic procedures for social research (revised). Newbury Park, CA: Sage.
- Rosenthal, R., & Rubin, D. (1978). Interpersonal expectancy effects: the first 345 studies. Behavior and Brain Sciences, 3, 377-415.
- Rosenthal, R., & Rubin, D. (1982). Comparing effect sizes of independent studies. Psychological Bulletin, 92, 500-504.
- Rosenthal, R., & Rubin, D. (1988). Comment: Assumptions and procedures in the file drawer problem. Statistical Science, 3, 120-125.
- Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. American Psychologist, 47 (10), 1173-1181.
- Schmidt, F. L., & Hunter, J. E. (1999). Comparison of three meta-analysis methods revisited: An analysis of Johnson, Mullen, and Salas (1995). Journal of Applied Psychology, 84 (1), 144-148.
- Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. Journal of Applied Psychology, 65, 643-661.
- Schmidt, F. L., Hunter, J. E., & Raju, N. S. (1988). Validity generalization and situational specificity: a second look at the 75% rule and the Fisher's Z transformation. Journal of Applied Psychology, 73, 665-672.
- Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: should Fisher's Z transformation be used? Journal of Applied Psychology, 72, 146-148.
- Strube, M. J. (1988). Averaging correlation coefficients: influence of heterogeneity and set size: Journal of Applied Psychology, 73, 559-568.

Wolf, F. M. (1986). Meta-Analysis. Sage university paper series on quantitative applications in the social sciences, 07-061. Newbury Park, CA: Sage.