British Journal of Mathematical and Statistical Psychology (2009), 62, 97–128

The British Psychological Society

www.bpsjournals.co.uk

Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results

Frank L. Schmidt[1]*, In-Sue Oh[1] and Theodore L. Hayes[2]

[1]Department of Management and Organizations, University of Iowa, Iowa City, USA
[2]Gallup Organization, Washington, DC, USA

Today most conclusions about cumulative knowledge in psychology are based on meta-analysis. We first present an examination of the important statistical differences between fixed-effects (FE) and random-effects (RE) models in meta-analysis and between two different RE procedures, due to Hedges and Vevea, and to Hunter and Schmidt. The implications of these differences for the appropriate interpretation of published meta-analyses are explored by applying the two RE procedures to 68 meta-analyses from five large meta-analytic studies previously published in *Psychological Bulletin*. Under the assumption that the goal of research is generalizable knowledge, results indicated that the published FE confidence intervals (CIs) around mean effect sizes were on average 52% narrower than their actual width, with similar results being produced by the two RE procedures. These nominal 95% FE CIs were found to be on average 56% CIs. Because most meta-analyses in the literature use FE models, these findings suggest that the precision of meta-analysis findings in the literature has often been substantially overstated, with important consequences for research and practice.

1. Introduction

In psychology, medicine, and the social sciences, conclusions about cumulative knowledge today are typically based on the results of meta-analyses. One indication of this is the large number of meta-analyses appearing in research journals in psychology and related areas, including journals that formerly published only individual empirical studies. Another indication is that textbooks summarizing knowledge within fields increasingly cite meta-analyses rather than a selection of primary studies, as was the case until recently (Hunter & Schmidt, 1996; Myers, 1991). Because conclusions about

* *Correspondence should be addressed to Dr Frank L. Schmidt, Department of Management and Organizations, Henry B. Tippie College of Business, University of Iowa, Iowa City, IA 52242-1994, USA (e-mail: frank-schmidt@uiowa.edu).*

DOI:10.1348/000711007X255327

cumulative knowledge are dependent on the meta-analysis methods used, it is important to examine carefully the implications of different statistical approaches to meta-analysis.

An important distinction within meta-analysis methods is that between fixed-effects (FE) and random-effects (RE) models. The differences between these two models have been discussed by Becker and Schram (1994), Field (2003), Hedges and Vevea (1998), Hunter and Schmidt (2000), the National Research Council (1992), Overton (1998), Raudenbush (1994), Shadish and Haddock (1994), and Schulze (2004). The basic distinction here is that FE models assume *a priori* that exactly the same population value (for example, ρ when *r* is the statistic used and δ when *d* is the statistic used) underlies all studies in the meta-analysis (i.e. $SD_\rho = 0$ or $SD_\delta = 0$), while RE models allow for the possibility that population parameters (ρ or δ values) vary from study to study.

The RE model is the more general one: the FE model is a special case of the RE model in which $SD_\delta = 0$. Application of an RE model can result in an estimated SD_δ (or SD_ρ) of zero, a finding indicating that an FE model would be appropriate for that set of studies. The application of an RE model can detect the fact that $SD_\delta = 0$; however, the application of an FE model cannot estimate SD_δ if $SD_\delta > 0$. The RE model allows for any possible value of SD_δ, while the FE model allows only the assumed value of zero. These differences in assumptions lead to different formulas for the standard error of the mean *d* or mean *r* which then lead to differences in the widths of estimated confidence intervals (CIs). In this paper we first present a discussion of the general statistical differences between FE and RE models and a discussion of the considerations surrounding their use. Next, we present a tabulation of their frequency of use in *Psychological Bulletin*, the premier US psychology review journal. We then discuss some technical questions in the estimation of RE models. Finally, we present reanalyses of data from five FE meta-analysis studies (68 meta-analyses in all) previously published in *Psychological Bulletin*, illustrating via empirical data that results and conclusions in psychological research – in particular, conclusions about the certainty of findings – depend importantly on which model is used. To our knowledge, no such demonstration based on reanalysis of archival data has appeared in the literature.

2. Differences between the two models

In psychology, the statistic averaged across studies is usually the correlation coefficient (*r*) or the standardized difference between means (the *d* statistic). The computed standard error (*SE*) of the mean *d* or *r* is a function of sampling error in the mean. There are two sources of sampling error: simple sampling error, estimated by the sampling error variance formula for *d* or *r*; and sampling error variance created by variation across studies in the underlying population values (i.e. S^2_δ or S^2_ρ). FE models consider only the first source of sampling error and do not take into account the second source (Field, 2001, 2003, 2005; Hedges & Vevea, 1998; Hunter & Schmidt, 2000; Overton, 1998; Raudenbush, 1994; Schulze, 2004). Hence the estimate of sampling error variance for the meta-analysis mean is accurate in the FE model only when $S^2_\delta = 0$ or $S^2_\rho = 0$. Otherwise, FE models underestimate sampling error variance and hence underestimate the *SE* of the mean *d* or *r*, leading to CIs that are too narrow (and also to inflated Type I error rates; Field, 2001, 2003, 2005; Hedges & Vevea, 1998; Hunter & Schmidt, 2000; National Research Council, 1992; Raudenbush, 1994; Schulze, 2004).

The most commonly used FE procedure is that of Hedges and Olkin (1985). In this procedure, the simple sampling error variance (V_{e_i}) is first computed for each study and the

inverses of these values ($1/V_{e_i}$) are the weights (w_i) used to compute the mean d value:

$$\hat{\delta} = \bar{d} = \frac{\sum w_i d_i}{\sum w_i}. \tag{1}$$

The sampling error variance of this mean ($S^2_{e_{\bar{d}}}$) – or rather, its square root, $SE_{\bar{d}}$ – is used to compute the CI in the usual manner; if a significance test is applied to the mean, it is also based on this *SE*. If the statistic used is *r*, Hedges and Olkin (1985) first transform the correlations using Fisher's z transformation; the calculations are carried out in Fisher's z metric, and then the mean and the endpoints of the CIs are back-transformed to the *r* metric (Hedges & Olkin, 1985, p. 120.) The FE procedure of Rosenthal and Rubin (Rosenthal, 1991, 1993; Rosenthal & Rubin, 1982a, 1982b) differs in only minor ways (Field, 2005).

If the study effect sizes are in the *d* metric, the simple sampling error variance for a single study is estimated as

$$V_{e_i} = \frac{N_1 + N_2}{N_1 N_2} + \frac{d^2}{2(N_1 + N_2)}, \tag{2}$$

where N_1 and N_2 are the sample sizes in the two groups (e.g. experimental and control groups) being compared and d is the standardized mean difference between the two groups (see Hedges & Olkin, 1985, equation 5–14, p. 86; Hunter & Schmidt, 2004, equation 7.23a, p. 284). The *d* statistic is usually corrected for its slight positive bias due to small sample size (see Hedges & Olkin, 1985, p. 81; Hunter & Schmidt, 2004, pp. 284–285). (When *r* is the statistic used, it is corrected for its slight negative bias before being transformed into Fisher's z metric.)

The reader can most clearly grasp the mechanics of the FE method in the special case in which the sampling error variance is constant across the studies in the meta-analysis. In this special case, the average of these values across studies is

$$\bar{V}_e = \frac{\sum V_{e_i}}{\boldsymbol{k}}. \tag{3}$$

The sampling error variance of the mean and the standard error are given by

$$S^2_{e_{\bar{d}}} = \frac{\bar{V}_e}{\boldsymbol{k}} \tag{4}$$

and

$$SE_{\bar{d}} = \sqrt{\frac{\bar{V}_e}{\boldsymbol{k}}}. \tag{5}$$

When sampling error varies across studies, the equation for sampling error variance becomes

$$S^2_{e_{\bar{d}}} = \frac{1}{\sum w_i}, \tag{6}$$

where the w_i are $1/V_{e_i}$. The *SE* is used with the mean d to compute the CI in the usual manner (Hedges & Olkin, 1985; Hedges & Vevea, 1998).

In the RE model, by contrast, the sampling error variance of the mean and that of the standard error are (again, assuming equal V_{e_i} across studies)

$$S^2_{e_{\bar{d}}} = \frac{\bar{V}_e}{\boldsymbol{k}} + \frac{S^2_{\delta}}{\boldsymbol{k}} \tag{7}$$

and

$$SE_{\bar{d}} = \sqrt{\frac{\bar{V}_e + S_\delta^2}{k}}. \tag{8}$$

The second term on the right-hand side of equation (7) reflects the effect of variance across studies in population parameters on the sampling error variance of the mean observed d. This term does not appear in the formula used by the FE model because the FE model assumes S_δ^2 (or SD_δ) to be zero; that is, the FE model assumes that the population parameters (ρ_i or δ_i) underlying each study are equal. As seen here, procedures for estimating sampling error variance for FE models are quite simple. These procedures are more complex for RE models and are presented later in this paper.

The methods described in Hunter, Schmidt, and Jackson (1982), Hunter and Schmidt (1990, 2004), Callender and Osburn (1980), and Raju and Burke (1983) are RE models (Hedges & Olkin, 1985, Ch. 9, p. 242; National Research Council, 1992, pp. 94–95). These methods have been extensively applied to substantive questions in the published literature (e.g. see Hunter & Schmidt, 1990, 2004; Schmidt, 1992). (These methods take into account artefacts in addition to sampling error, such as measurement error.) The methods described in articles by Hedges (1983, 1988), Hedges and Olkin (1985, Ch. 9), Raudenbush (1994), Raudenbush and Bryk (1985), and Rubin (1980, 1981) are also RE methods. These latter methods have been used less frequently in meta-analysis. For example, although *Psychological Bulletin*, the major review journal in psychology, has published 199 meta-analyses as of January 2006, we could locate only 13 meta-analyses published in that journal that employed these methods. (See discussion in later section.) Schulze (2004, p. 35) noted that the FE model has been more commonly used than the RE model, and Cooper (1997, p. 179) stated that, 'In practice, most meta-analysts opt for the fixed effects assumption because it is analytically easier to manage'. The National Research Council (1992, p. 52) stated that many users of meta-analysis prefer FE models because of 'their conceptual and computational simplicity'.

An important question is whether the FE assumption of constant population parameter values can accurately reflect reality. Many would argue that for theoretical or substantive reasons there is always some variation in population parameter values across studies (National Research Council, 1992; Schulze, 2004). That is, they would argue that there are always at least some real (i.e. substantive, not methodological) moderator variables (interactions) that create differences in values of δ_i or ρ_i across studies. However, evidence has been reported indicating that some study domains appear to be homogeneous at the level of substantive population parameters (Hunter & Schmidt, 2000; Schmidt *et al.*, 1993). That is, population parameters do not vary once the effects of sampling error, measurement error, and range variation are controlled for. However, such homogeneity can be revealed only by using RE models to estimate the level of heterogeneity. FE models do not allow for such calibration because they assume homogeneity *a priori*.

Even if there are no substantive moderators causing variation in population parameters, methodological variations across studies can cause variation in study population δ_i or ρ_i values; that is, values corresponding to $N = \infty$ can be affected by methodological variations other than sampling error. For example, if the amount of measurement error (degree of unreliability) in the measures used varies across studies, then this variation creates variation in study population parameters; studies with more measurement error will have smaller study population values of δ_i or ρ_i, and vice versa

(Hedges & Olkin, 1985, pp. 135–138; Hunter & Schmidt, 2004, Ch. 3). So even if there is no substantive variation in population parameters, variations across studies in such methodological factors as error of measurement, range variation, or dichotomization of continuous variables (Hunter & Schmidt, 1990, 2004; Osburn & Callender, 1992) will create variation in study population parameters. In the absence of corrections for such artefacts such variation will typically exist, causing the assumption of homogeneous study population effect sizes or correlations to be false and the CIs based on the FE model to be too narrow.

Hedges and Vevea (1998) and Overton (1998) pointed out that the choice of an FE or RE model depends on the type of inference that is the goal of the meta-analyst. If the goal is to draw conclusions that are limited to the set of studies at hand and the meta-analyst does not desire to generalize beyond his/her particular set of studies, the FE model can be used when population parameters vary as well as when they do not. Hedges and Vevea refer to this as conditional inference. The usual goal of research, however, is generalizable knowledge (Toulmin, 1961), which requires generalization beyond the current set of studies to other similar studies that have been or might be conducted. Hedges and Vevea refer to this as unconditional inference. Within this broader objective, the FE model is appropriate only when population parameters do not vary. When population parameters vary, an RE model is required for unconditional inference (Field, 2003, 2005; Hedges & Vevea, 1998; Raudenbush, 1994). The discussion in this paper assumes that the objective in meta-analysis is to make inferences about a wider population of studies; that is, to draw conclusions that can be generalized beyond the specific set of studies included in the meta-analysis. If this is not the case and the researcher's purpose is only to reach conclusions limited to the specific set of studies in the meta-analysis, the FE model does not underestimate the *SE* and the resulting CIs are not too narrow. This follows from the fact that in this case there is no sampling error in the sampling of study population parameters, because the set of studies at hand is not viewed as a sample of a larger number of studies that might exist or could be conducted (Hedges & Vevea, 1998; Hunter & Schmidt, 2000; Overton, 1998; Raudenbush, 1994). In this case, generalization of conclusions is only to a hypothetical set of studies that is identical to the study set at hand except for simple sampling error; that is, to a set of studies with exactly the same study population parameter values, study for study, and differing only in the sampling of subjects (usually people) within studies.

Schulze (2004, pp. 38, 195) stated that it is difficult for a meta-analyst to decide whether his/her purpose is this limited generalization and also difficult for a reader of the meta-analysis to evaluate such a decision and that this creates difficulties in interpreting FE results when $S_\delta^2 > 0$ or $S_\rho^2 > 0$. More importantly, it is has been pointed out that such conclusions are of limited scientific value (Hedges & Vevea, 1998; Hunter & Schmidt, 2000; National Research Council, 1992; Overton, 1998; Schulze, 2004). The goal of science is cumulative knowledge, and cumulative knowledge is generalizable knowledge (Bechtel, 1988; Phillips, 1987; Toulmin, 1961). Researchers are interested in general principles, not in describing a particular set of studies. Hence, it would appear that the FE model would rarely be appropriate for most research purposes. The National Research Council (1992) stated that FE models 'tend to understate actual uncertainty' (p. 147) in research findings and recommended 'an increase in the use of random effects models in preference to the current default of fixed effects models' (p. 2; see also pp. 185–187 of that report). Others have also cautioned that when the goal is generalizable knowledge, use of FE models can lead to inflated Type I error rates and erroneously narrow confidence intervals (e.g. Field, 2003; Hedges, 1994; Hedges & Vevea, 1998; Hunter & Schmidt, 2000; Overton, 1998; Raudenbush, 1994; Rosenthal, 1991).

Hedges and Vevea (1998, pp. 487–488) stated that although there is no statistical (sampling) foundation or justification for generalizing FE findings beyond the specific studies in the meta-analysis, there can be, by analogy with the practices of some primary researchers using analysis of variance (ANOVA) in experiments, an extra-statistical or judgement-based basis for such wider generalization. They proposed that just as primary researchers using FE ANOVA designs in experiments sometimes generalize their conclusions beyond the specific fixed levels of treatment included in their experiments, so also could meta-analysts using FE models, based on the subjective judgement that new studies will be 'sufficiently similar' to those in the meta-analysis. In ANOVA, an FE design is one in which all levels of the treatment that are of interest are included in the design, while an RE model is one in which only a sample of treatment levels of interest is included in the study. It was by analogy with this distinction in ANOVA that Hedges and Olkin (1985, p. 149) originally labelled the two different models in meta-analysis as FE and RE models (Hedges & Vevea, 1998). Hence in FE meta-analysis models, the studies included in the meta-analysis are assumed to constitute the entire universe of relevant studies, whereas in RE models the studies are taken to be a sample of all possible studies that might be conducted or might exist on the subject. However, the National Research Council report (1992, pp. 46 and 139) indicates that there are problems with this analogy:

> The manner in which the terms 'fixed effects' and 'random effects' are used in the meta-analysis literature is somewhat different from the classical definitions used in other techniques of statistics such as analysis of variance, where 'fixed effects' is the term required to deny the concept of a distribution of the true effects, $\delta_1 \ldots \delta_k$, and 'random effects' supposes that the δ_i are sampled from a population and therefore have a distribution. (National Research Council, 1992, p. 46)

An example might help to clarify the meaning of this National Research Council statement. A study of the effects of a drug on patients might include the dosages 0, 10, and 20 mg. In FE ANOVA, treatments (dosages) are fixed at these levels and these levels are considered the only ones of interest. In the FE ANOVA the idea that there is a naturally occurring distribution of dosages or potential dosages is explicitly denied. This is different from the FE model in meta-analysis in two ways. First, in meta-analysis, the researcher does not specify (or fix) the parameter values (ρ_i or δ_i) in the individual studies included in the meta-analysis; instead, these values are accepted as they happen to be sampled in the set of studies at hand. That is, they are observed and not manipulated. The second difference results from the first: because the researcher does not fix the parameter values included in the studies but accepts them as they happen to have occurred, there appears to be no basis or rationale for postulating or assuming that these parameter values do not have a distribution across studies, which is the key assumption of the FE model in ANOVA. This is the reason why the National Research Council (1992) report rejected the analogy between FE models in ANOVA and FE models in meta-analysis. However, even had the National Research Council accepted this analogy at the conceptual level, this would still have left open the question of whether the broader generalizations sometimes made by researchers from FE ANOVA-based experiments are justified. The fact that experimenters sometimes make such generalizations cannot be viewed as a justification (Schulze, 2004). As Hedges and Vevea (1998) pointed out, this practice has no statistical foundation and is based only on subjective judgement. The National Research Council (1992) report concluded that unless population parameters actually do not vary, FE models will yield CIs that are too narrow (and inflated Type I error rates) when there is any generalization to studies beyond the specific ones included in the meta-analysis. This is also the conclusion

of Field (2001, 2003, 2005), Hunter and Schmidt (2000), Overton (1998), Schulze (2004), and others.

Potential conceptual problems are also associated with the use of the RE model. In that model, the studies in the meta-analysis are viewed as a random sample from a larger universe of studies that exist or could be conducted. Hedges and Vevea pointed out that this larger universe is often poorly defined and ambiguous in nature. However, Schulze (2004, pp. 40–41) noted that this is not a problem specific to meta-analysis or RE models in meta-analysis, but one that characterizes virtually all samples used in research. Rarely in research is the target population of samples fully enumerated and delimited; in fact, data sets used frequently consist of something close to convenience samples (i.e. a set of samples for which it was possible to obtain data). Viewed in this light this problem appears less serious. Another potential problem with RE models is the fact that in the estimation of the between-study parameter variance (S_δ^2 or S_ρ^2), the number of data points is the number of studies. Hence, if the number of studies is small, estimation of this quantity can have less than sterling accuracy (Hedges & Vevea, 1998; Hunter & Schmidt, 1990, 2004; Raudenbush, 1994). One implication is that results produced by RE models should be considered only approximate when the number of studies is small. In the reanalysis reported later in this paper, we include data only for meta-analyses with 10 or more studies.

To the extent that empirical results and conclusions differ depending on whether FE or RE models are used, it is important, as Overton (1998) pointed out, to examine these differences and determine the extent to which conclusions about cumulative knowledge depend on assumptions underlying the two models. Overton (1998, p. 357) suggested that data should be analysed using both models to reveal the extent to which findings are dependent on the specific assumptions of the FE and RE models.

3. Use of FE and RE models in psychological literature

At present most meta-analyses in the psychological literature appear to be based on FE methods (Cooper, 1997, p. 179; Schulze, 2004, pp. 35, 82–83). The National Research Council (1992, p. 146) stated that the use of FE models is 'the rule rather than the exception'. To provide some empirical calibration, we tabulated the frequency of use of FE and RE models in *Psychological Bulletin*, from the first appearance of a meta-analysis in 1978 to the January 2006 journal issue; the results are shown in Table 1. During this period 199 meta-analysis studies were published. Because it does not address the question of sampling error, the Glass procedure for meta-analysis (Glass, McGaw, & Smith, 1981) cannot be classified as an FE or RE procedure. The 10% of studies using this approach are concentrated in the first half of this period; after 1994, no studies used this procedure. Another 6% of meta-analysis studies also used procedures that were too rudimentary to be classified as FE or RE methods. Of the 169 meta-analysis studies that could be classified, 129 (76%) used only FE methods. Of the 129 FE meta-analysis studies, 91 (71%) employed the FE procedures of Hedges and colleagues (e.g. Hedges & Olkin, 1985), 24 (19%) used the Rosenthal and Rubin FE procedure, and 14 (11%) used combinations or did not provide enough information to allow classification.

Of the 32 meta-analysis studies that used RE models, 8 (25%) used the RE procedures of Hedges and colleagues (e.g. Hedges & Olkin, 1985; Hedges & Vevea, 1998), 19 (59%) used the Hunter–Schmidt RE procedure (Hunter *et al.*, 1982; Hunter & Schmidt, 1990, 2004), and 5 (16%) used other RE procedures. Of the eight meta-analysis studies that used both FE and RE models, seven (88%) used Hedges and colleagues' procedures

Table 1. Classification of meta-analysis articles in *Psychological Bulletin*, 1977–2006

	Methods							
Year	Glassian	FE	RE	Both (mixed)	Misc.	Percentage of FE[a]	Sum	Cumulative sum
1977							0	0
1978					1		1	1
1979					1		1	2
1980							0	2
1981	1	3			1	100	5	7
1982	3	1				100	4	11
1983	2	2				100	4	15
1984	1	2	1		1	67	5	20
1985		6	2			75	8	28
1986	4	6	1			86	11	39
1987	2	3	1			75	6	45
1988	1	5	1			83	7	52
1989	1	5				100	6	58
1990	2	5	2			71	9	67
1991	1	7	1			88	9	76
1992		9				100	9	85
1993		5	1			83	6	91
1994	1	6				100	7	98
1995		8	1			89	9	107
1996		8	2		2	80	12	119
1997		5				100	5	124
1998		6	2			75	8	132
1999		8				100	8	140
2000		8	1	1	1	89	11	151
2001		2	2			50	4	155
2002		9			2	100	11	166
2003		6	3	3	1	67	13	179
2004		3	6	1		33	10	189
2005		1	3	2	1	25	7	196
2006			2	1		0	3	199
Total	19	129	32	8	11	76	199	
%	10	65	16	4	6		100	

Note. 2006 data include only the first issue or 132(1); Glassian: Glass and colleagues procedure, 19 (100%); FE: Hedges and colleagues procedure, 91 (71%), Rosenthal and colleagues procedure, 24 (19%), and other procedures, 14 (11%); RE: Hunter and Schmidt procedure, 19 (59%), Hedges and colleagues procedure, 8 (25%), and other procedures, 5 (16%); Both: Hedges and colleagues procedure, 7 (88%), and other, 1(13%); Misc.: unclassifiable meta-analysis procedures.

[a] Computed based on studies classified as FE, RE, and both. Studies classified as 'both' were counted as RE studies.

(e.g. Hedges & Olkin, 1985; Hedges & Vevea, 1998), and one (13%) used another method. None used the Rubin (1980, 1981) RE model.

Table 1 also shows for each year the percentage of meta-analysis studies that used the FE model. This column is based on the 169 studies that could be classified as FE, RE, or both, with the latter being considered RE studies. It can be seen that since about 2003 there has been some increase in the number of RE-based meta-analysis studies. Additional insight is

provided by Figure 1, which is also based on the 169 meta-analysis studies that used the FE model, the RE model, or both, with the latter again considered to be RE studies. (Many of the studies that used both models relied only on the FE results in their interpretations.) Figure 1 suggests that the proportion of studies using the FE model exclusively appears to be declining with time. This trend is likely due to the effect of the National Research Council (1992) report and to publications by Becker and Schram (1994), Hedges and Vevea (1998), Hunter and Schmidt (2000), Overton (1998), and Raudenbush (1994), among others. We view this trend as positive, but the fact still remains that over three quarters of classifiable meta-analyses published in *Psychological Bulletin* employed only the FE model.

4. Estimation in RE models

Procedures for estimation of sampling error variance are more complex in RE models than FE models (Schulze, 2004). In our reanalyses of published FE meta-analyses, we present results for both the Hedges–Vevea (HV) RE procedure and the Hunter–Schmidt (HS) RE procedure. Estimation procedures differ somewhat for these two RE procedures (cf. Field, 2005), although they generally yield similar results (and were found to do so in our analyses). Estimation procedures are simpler for the HS approach (Schmidt, Hunter, & Raju, 1988), so we present those procedures first.

4.1. The Hunter–Schmidt RE procedure

Again our presentation is in terms of the *d* statistic but procedures are similar and analogous for *r* (the correlation coefficient) and other indices of effect size. Here we

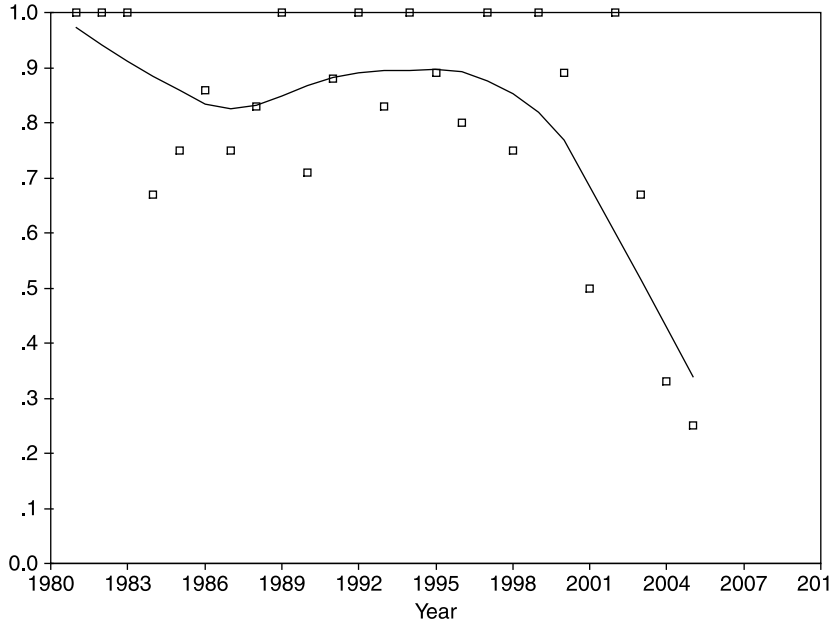

Figure 1. Proportion of FE-based meta-analysis articles in *Psychological Bulletin*, 1981–2005. Computed based on studies classified as FE, RE, and both ($k = 169$) from Table 1. Studies using both FE and RE models were counted as RE studies.

present only the basics of the HS RE method; a more technically detailed description can be found in (for example) Field (2005). In the HS RE procedure, the sampling error variance of the mean d is estimated as the variance of the observed ds across studies divided by k, the number of studies:

$$S^2_{e_{\bar{d}}} = \frac{\bar{V}_e}{k} + \frac{S^2_\delta}{k} = \frac{S^2_d}{k}. \tag{9}$$

The square root of equation (9) is the *SE* that is used in computing CIs:

$$SE_{\bar{d}} = \frac{SD_d}{\sqrt{k}} = \sqrt{\frac{\bar{V}_e + S^2_\delta}{k}}. \tag{10}$$

In this model, $\bar{V}_e$ is conceptualized as the sample size weighted mean of the V_{e_i} values. The equation for S^2_d is

$$S^2_d = \frac{\sum N_i(d_i - \bar{d})^2}{\sum N_i}, \tag{11}$$

where

$$\bar{d} = \frac{\sum N_i d_i}{\sum N_i}. \tag{12}$$

The rationale for this procedure can be seen in the fact that $S^2_d = S^2_e + S^2_\delta$; that is, the expected value of S^2_d is the sum of simple sampling error variance and the variance of the study population parameters (Field, 2005; Hedges, 1989; Hunter *et al.*, 1982; Hunter & Schmidt, 1990, 2004; Schmidt & Hunter, 1977).[1] Hence S^2_d estimates the average RE sampling error variance for the set of studies, and this quantity divided by k is the sampling error variance of the mean. Osburn and Callender (1992) showed that this equation holds both when $S^2_\delta > 0$ and $S^2_\delta = 0$ (i.e. when the assumption underlying the FE model holds). The study weights in the HS RE model are (total) study sample sizes, N_i, used because these weights closely approximate the inverse of the simple sampling error variances, $1/V_{e_i}$ (Hunter & Schmidt, 2004), and are less affected by sampling error variance (Brannick, 2006). Hedges (1983, p. 392) stated that in the heterogeneous case ($S^2_\delta > 0$), weighting by sample size 'will give a simple unbiased estimator [of the mean] that is slightly less efficient than the optimal weighted estimator'. Osburn and Callender (1992) showed via simulation that weighting by sample size produces accurate *SE* estimates both when $S^2_\delta = 0$ and $S^2_\delta > 0$, as long as multiple outlier sample sizes are not present. (In the presence of very large outlier sample sizes, weighting by sample size can cause underestimation of the *SE*.) Also using simulation, Schulze (2004) found that for heterogeneous population data sets, the HS RE procedure weighting by sample size produced accurate (more accurate than other procedures evaluated) estimates of CIs (Table 8.13, p. 156); estimates for the mean correlation were also acceptably accurate (with a small median negative bias of 0.0022; Table 8.4, p. 134; see pp. 188–190 for a summary). Brannick reported similar results. Further details can be found in Osburn and

[1] *The focus here is on the d statistic, but the general principle also applies to the correlation coefficient. In the case of the correlation coefficient, there is a non-zero covariance between population correlations and sampling errors. Hedges (1989) examined this issue and concluded that this covariance is very small and has only trivial effects on conclusions reached using this additive model. A similar conclusion applies to the d statistic.*

Callender (1992) and Schmidt *et al.* (1988). We note here that in the HS RE method, when the *d*s are corrected for measurement error, the procedure is analogous except that S_d^2 is now the variance ($S_{d_c}^2$) of the corrected *d*s (see Hunter and Schmidt, 2004, pp. 206–207 for details). We do not address such corrections in this paper.

4.2. The Hedges–Vevea RE procedure

The Hedges and Vevea (1998) RE procedure estimates the two components of RE sampling error variance separately. The simple sampling error variance component is estimated exactly as it is in the FE model. That is, equation (6), reproduced below, is used to compute this component:

$$S_{e_{\bar{d}}}^2 = \frac{1}{\sum w_i}$$

where the w_i are $1/V_{e_i}$.

The second component, $\hat{\sigma}_\delta^2$ (symbolized as $\hat{\tau}^2$ by Hedges and Vevea), is estimated as follows:

$$\hat{\sigma}_\delta^2 = \begin{cases} \frac{Q-(k-1)}{c}, & \text{if } Q \geq k-1, \\ 0, & \text{if } Q < k-1, \end{cases} \tag{13}$$

where Q represents the χ^2 overall homogeneity test[2] and c is a function of the study weights and is given in equation (11) from Hedges and Vevea (1998):

$$c = \sum w_i - \frac{\sum (w_i)^2}{\sum wi}, \tag{14}$$

where the study weights w_i are the FE study weights as defined in our equation (6).

The estimated mean value is then

$$\hat{\delta} = \bar{d} = \frac{\sum w_i^* d_i}{\sum w_i^*}. \tag{15}$$

The sampling error variance is:

$$S_{e_{\bar{d}}}^2 = \frac{1}{\sum w_i^*} \tag{16}$$

where the w_i^* are $1/(V_{e_i} + \hat{\sigma}_\delta^2)$.

When the effect size statistic is the correlation, this RE procedure first converts *r*s to the Fisher z transformation, conducts the calculations in that metric, and then back-transforms the resulting means and CIs into the r metric (Hedges & Olkin, 1985) (as is done in the FE procedure). See Hedges and Vevea (1998), Field (2005), and Hall and Brannick (2002) for a complete technical description of this RE procedure.

When $Q - (k - 1)$ yields a negative value, $\hat{\sigma}_\delta^2$ in equation (13) is set to zero because, by definition, a variance cannot be negative. Hedges and Vevea discuss in some detail the positive bias that characterizes this estimate as a result of setting negative values to zero,

[2] *It should be noted that a value of mean d is needed initially to compute the Q statistic. In computing that initial mean d, studies are weighted by w, not w*. That is, they are weighted by the inverse of the simple sampling error variance.*

and they tabulate this bias in their Table 2 for various conditions. This bias causes the *SE* to be upwardly biased, causing the resulting CIs to be too wide; that is, the probability content of the CIs is larger than the nominal value (Hedges & Vevea, 1998, p. 496). Overton (1998, pp. 371, 374) found this same bias for this procedure and also for an iterative procedure he used to estimate S^2_ρ and S^2_δ. Hedges and Vevea (1998, p. 492) briefly discussed such iterative procedures but rejected them in favour of the procedure described above because the iterative procedures are more complex to use, typically requiring specialized computer programs, and do not appear to be more accurate. Hedges and Vevea state that bias becomes smaller as *k* (the number of studies) increases and is generally small when *k* is 20 or more. However, Overton (1998) pointed out that the bias also depends on the actual size of S^2_δ (or S^2_ρ). For example, if this value is zero, then 50% of the estimates are expected to be negative due to sampling error, creating a positive bias regardless of the number of studies. If this value is small but not zero, then less than 50% of the estimates of S^2_δ are expected to be negative, and the positive bias is smaller. When S^2_δ is large, the positive bias is negligible. Overton (1998) stated that when S^2_δ is small, the RE model overestimates sampling error variance and produces CIs that are too wide. Some researchers have cited Overton's statement as a rationale for preferring the FE model to the RE model in their meta-analyses (e.g. Bettencourt, Talley, Benjamin, & Valentine, 2006).

Because of its different mode of estimating the sampling error variance (described above), the HS RE procedure does not have this upward bias. As shown earlier, in the HS RE procedure, the two components of the RE sampling error variance are estimated jointly rather than separately. Note that if S^2_δ is in fact zero, the HS RE estimate of sampling error variance has the same expected value as the FE estimate of sampling error variance (Osburn & Callender, 1992; Schmidt *et al.*, 1988). As shown by Hedges and Vevea (1998), this is not the case for the HV RE procedure.

Because of the nature of the study weights used to produce the weighted mean *d* value in the HV procedure, it is necessary to have a separate estimate of σ^2_δ when using this procedure (Field, 2005; Hedges & Vevea, 1998). As noted above, the weight applied to each study is $w_i^* = 1/(V_{e_i} + \hat{\sigma}^2_\delta)$, where V_{e_i} is the simple sampling error variance for that study. The HS procedure weights each study by its (total) sample size (N_i) and therefore does not require a separate estimate of σ^2_δ. The HS RE model does estimate σ^2_δ for other purposes (cf. Hunter & Schmidt, 2004), and this estimate does have a positive bias, but this estimate is not used in computing weighted mean values, *SE*s, or CIs (Schmidt *et al.*, 1988; see also Schulze, 2004, p. 190). Within the context of large-sample statistical theory, the HV study weights are statistically more appropriate for RE models (Hedges, 1983; Hedges & Vevea, 1998; Raudenbush, 1994; Schulze, 2004), but even within large-sample theory this advantage is slight (Hedges, 1983, p. 393). In addition, the small theoretically expected advantage for these study weights is not realized with the smaller study sizes that are typical, because of inaccuracies induced by sampling error in the estimation of the σ^2_δ component of the weights (e.g. see Brannick, 2006; Raudenbush, 1994, p. 317; and Schulze, 2004, pp. 84 and 184). Because of this effect, Schulze (2004, pp. 193–194), based on the results of his extensive Monte Carlo studies, recommended weighting studies by sample size in the heterogeneous case (i.e. σ^2_δ or $\sigma^2_\rho > 0$), as well as the homogeneous case.

In a recent simulation study, Field reported that the HS RE model yielded generally more accurate estimates of mean values than the HV RE model, but this research was limited to correlations and it is possible that its findings may be explained by a positive bias induced in estimates of mean *r* in the HV model by its use of the non-linear Fisher *z* transformation of correlations (Donnor & Rosner, 1980; Field, 2001, 2005; Hotelling,

1953; Hunter, Schmidt, & Coggin, 1996; Overton, 1998, p. 358; Schulze, 2004, pp. 75–79, 193–194). Hall and Brannick (2002) reported findings similar to those of Field (2005). This bias becomes larger as S_p^2 increases and can become substantial. In the Field (2001) simulation study, this bias was as large as 0.20; in Field (2005), which employed a slightly different simulation method, the maximum bias was 0.052. The bias reported by Hall and Brannick (2002) was intermediate in value. In the units of the *d* statistic, these bases would be slightly more than twice their value in the *r* metric (Hunter & Schmidt, 2004). Schulze (2004), also using the *r* statistic, reported this same finding in his Monte Carlo studies and attributed it to use of the Fisher *z* transformation by the HV RE method. He recommended against use of this transformation (Schulze, 2004, pp. 193–194). No comparable transformation is used in the HV RE method when the *d* statistic is used in the meta-analysis. The meta-analyses that were re-analysed in the present study were limited to those using the *d* statistic so that the results produced by the HV and HS RE procedures would be more directly comparable and could be interpreted without the distracting issue of the bias caused by the Fisher *z* transformation. However, as discussed later, the general pattern of results can be expected to be the same for *r* as for *d*.

5. Method

We sought to reanalyse data from FE meta-analyses published in *Psychological Bulletin*. We searched for studies that met the following criteria: (a) the FE model only was used; (b) CIs were presented and interpreted; (c) data tables presented effect sizes, *N*s, and information needed to code the studies into the categories used in the meta-analysis; and (d) the study used the *d* statistic. Surprisingly, few studies met requirement (c), limiting our choice of meta-analysis studies. We found only four meta-analysis studies that met all four of the criteria. We believe these meta-analysis studies are methodologically typical of those that have appeared in *Psychological Bulletin* over the last 20 years, except for the fact that they provided the data necessary to replicate (recompute) the meta-analyses. That is, these studies are typical in their use of and interpretation of the FE method. We believe that this – not substantive (topic area) typicality – is what is important for our purposes, which are methodological in nature. Chronologically, these studies were: (a) Hyde and Linn (1988), a study of gender differences in verbal ability; (b) Hyde, Fennema, and Lamon (1990), a study of gender differences in mathematics performance; (c) Bettencourt and Miller (1996), a study of gender differences in aggression; and (d) Byrnes, Miller, and Schafer (1999), a study of gender differences in risk taking. A reviewer requested the inclusion of at least one meta-analysis focusing on a substantive area other than gender differences. After considerable search we located the Bettencourt *et al.* (2006) study, a meta-analysis focusing on relations between personality and aggressive behaviour under provoking and neutral conditions. This study reported both FE and RE results (and so violated our condition (a) above), but the authors based all their interpretations of results on only the FE results. We could find no other studies that met requirement (b), (c), and (d). All five of these studies reported multiple meta-analyses, with a total across studies of 68 separate meta-analyses with $k = 10$ or more. All employed the Hedges and Olkin (1985; Hedges & Vevea, 1998) FE meta-analysis procedure; all reported (nominal) 95% CIs. Since all meta-analyses become increasingly less accurate as *k* (the number of studies) becomes smaller (Hedges & Vevea, 1998), we limited our reanalysis to meta-analyses based on 10 or more studies.

5.1. Procedure

We wrote spreadsheet-based programs for the Hedges and Olkin FE and RE procedures and calibrated these programs against the example analysis given in Hedges and Vevea (1998; Table 1). Next, we re-analysed each of the 68 meta-analyses using the Hedges and Olkin (1985) FE procedure (also described in Hedges & Vevea, 1998; see also Field, 2005) and confirmed that results obtained were identical or nearly identical to the originally reported results. (In a few cases, we were unable to locate data for one or two of the studies originally reported to be in the meta-analysis, resulting in slightly different results. For example, authors might report that $k = 21$ for a meta-analysis of verbal ability but we could find only 20 studies of verbal ability in the data table presented.) In four of the studies (Hyde & Linn, 1988; Hyde *et al.*, 1990; Bettencourt & Miller, 1996; Bettencourt *et al.*, 2006), it was specified that the Hedges and Olkin (1985) adjustment for the slight positive bias in the *d* statistic had been applied to the *d* statistics presented in the data tables, and we used these adjusted values in our reanalysis. In the remaining meta-analysis (Byrnes *et al.*, 1999) it was not stated that this adjustment had been applied, so we applied it before our reanalysis. (The results were almost identical with and without this adjustment.) After the reanalysis using the HV FE method, we reanalysed these same data sets using the Hunter–Schmidt RE model described earlier (Field, 2005; Hunter & Schmidt, 2004; Schmidt *et al.*, 1988) and the Hedges–Vevea RE model (Field, 2005; Hedges & Vevea, 1998), also described earlier. Both these procedures include an adjustment for the slight positive bias in *d* values, and these adjustments were included (when needed). (Spreadsheet programs used are available from the first author.) For the FE and the two RE procedures, we computed means and 95% CIs following the usual procedures for computing CIs. Our major focus was on CIs because both Hedges and Olkin (1985) and Hunter and Schmidt (1990, 2004; see also Hunter *et al.*, 1982) recommend that CIs be presented in preference to statistical significance tests. All CIs were computed based on the normal (*z* score) distribution, because this distribution (rather than the *t* distribution) was used to compute CIs in the original studies. For both RE models, we computed the percentage by which the FE model underestimated the RE CI. We also computed the probability value (confidence level) of the FE CI using the RE CI as the standard. For example, a nominal 95% FE CI might actually be a 65% CI when evaluated against an RE model CI. These two indices provide informative measures of the differences between estimates of precision for FE and RE models.

6. Results

Results are presented in Tables 2–7. The table numbers within our tables indicate the tables in the original studies from which the specific meta-analyses were taken. In each of Tables 2–7, the first section gives the results of the FE analysis as reanalysed by us, presenting mean *d*s and nominal 95% FE CIs. The middle section shows results for the HS RE procedure, and the final section shows the results for the HV RE procedure. For both RE procedures, mean *d*s and CIs are presented followed by three additional columns of information. The column headed 'Diff' presents the amount by which the FE model underestimated the actual CI. The next column gives the percentage by which the FE model underestimated the actual CI. The next column gives the actual confidence levels of the nominally 95% FE CIs, showing that the FE CIs typically do not reach the 95% CI.

Table 2. Reanalyses of Tables 4 and 5 of Hyde and Linn (1988)

		Fixed-effects model			Random-effects models										
		Hedges–Olkin			Hunter–Schmidt						Hedges–Vevea					
Test type	*k*	$\hat{\hat{\delta}}$	95% CI		$\hat{\hat{\delta}}$	95% CI		Diff	% Under	% CI	$\hat{\hat{\delta}}$	95% CI		Diff	% Under	% CI
							Table 4									
Vocabulary	40	.02	−.02	.06	.02	−.05	.09	.06	43	74	.04	−.05	.12	.09	53	59
Reading comprehension	19[a]	.03	.01	.04	.03	−.01	.07	.05	63	52	.09	.03	.14	.08	73	4
Speech production	12	.33	.20	.46	.33	.20	.46	.00	0	95	.33	.20	.46	.00	0	95
General/mixed	25	.20	.19	.21	.21	.12	.29	.15	88	18	.22	.12	.33	.19	90	14
Total	120[b]	.11	.10	.12	.11	.09	.14	.03	60	57	.14	.10	.18	.06	75	14
							Table 5									
RD	29	.09[c]	.04	.13	.09	.01	.16	.06	40	76	.11	.03	.19	.07	44	64
RP	13	−.12	−.19	−.04	−.12	−.24	.00	.09	38	78	−.16	−.32	−.01	.16	52	58
RD + A + S	18[d]	.03[c]	.01	.04	.03	−.02	.07	.06	67	48	.09	.02	.15	.10	77	6
P	17	.10[c]	.09[c]	.11[c]	.10	.07	.12	.03	60	57	.36	.18	.53	.33	94	0
M	28	.19	.18	.20	.20	.12	.28	.14	88	19	.21	.11	.31	.18	90	14

Note. FE results are replicated results; % Under means the percentage underestimation comparing the FE CI to the RE CI; % CI means the true confidence level (%) comparing the FE CI to the RE CI; RD, retrieval of definition of a word; RP, retrieval of the name of a picture; A, analysis of relations among words; S, selection of relevant information; P, production (written or oral); M, mixture of processes.

[a] The *k* in the original was 18.

[b] The *k* in the original was 119.

[c] The replication of original results is off by .01, which can be attributed to rounding errors.

[d] The *k* in the original was 17.

Table 3. Reanalyses of Tables 2 and 3 of Hyde *et al.* (1990)

		Fixed-effects model			Random-effects models										
		Hedges–Olkin			Hunter–Schmidt						Hedges–Vevea				
Test type	*k*	$\hat{\bar{\delta}}$	95% CI		$\hat{\bar{\delta}}$	95% CI		Diff	% Under	% CI	$\hat{\bar{\delta}}$	95% CI		Diff	% Under	% CI
							Table 2									
Computations	45	−.14	−.14	−.13	−.14	−.18	−.10	.07	88	19	−.14	−.19	−.10	.08	89	17
Concepts	41	−.03	−.04	−.02	−.03	−.04	−.01	.01	33	81	−.04	−.06	−.01	.03	60	44
Problem solving	48	.09[a]	.07	.11[a]	.09	.02	.16	.10	71	42	.19	.11	.26	.11	73	2
Mixed or unreported	121[b]	.19	.19[a]	.19	.19	.15	.23	.08	100	0	.16	.11	.21	.10	100	0
							Table 3									
Arithmetic	35	.00	−.02	.01	.00	−.05	.04	.06	67	48	.03	−.02	.09	.08	73	20
Geometry	19	.12[a]	.08[a]	.15[a]	.12	.05	.18	.06	46	70	.10	.03	.18	.08	53	60
Mixed or unreported	190	.15	.15	.16[a]	.16	.12	.19	.06	86	21	.09	.05	.13	.07	88	0

Note. FE results are replicated results; % Under means the percentage underestimation comparing the FE CI to RE CI; % CI means the true confidence level (%) comparing the FE CI to the RE CI.

[a] The replication of original results is off by .01, which can be attributed to rounding errors.

[b] The *k* in the original was 121.

Table 4. Reanalyses of Tables 4, 6, and 7 of Bettencourt and Miller (1996)

		Fixed-effects model			Random-effects models											
		Hedges–Olkin			Hunter–Schmidt						Hedges–Vevea					
Category	*k*	$\hat{\bar{\delta}}$	95% CI		$\hat{\bar{\delta}}$	95% CI		Diff	% Under	% CI	$\hat{\bar{\delta}}$	95% CI		Diff	% Under	% CI
							Table 4									
Overall[a]	107	.23[b]	.18	.29	.25	.15	.35	.09	45	70	.24	.14	.33	.08	42	74
Overall	107	.24	.19	.29	.26	.17	.35	.08	44	68	.24	.14	.33	.09	47	70
Provocation[a]	66	.75[b]	.66	.84[b]	.89	.65	1.13	.30	63	31	.86	.64	1.09	.27	60	39
Provocation	66	.75	.66	.83	.87	.64	1.09	.28	62	33	.86	.64	1.07	.26	60	36
							Table 6									
Neutral condition	50	.33	.24[b]	.42	.37	.20	.53	.15	45	66	.39	.22	.55	.15	45	60
Provocation condition	57	.17	.09	.24[b]	.17	.05	.28	.08	35	80	.13	.02	.24	.07	32	74
							Table 7									
Provocation condition																
Insult	17[c]	−.10[d]	−.26[d]	.05[d]	−.13	−.35	.09	.13	30	82	−.12	−.32	.07	.08	21	88
Physical attack	13	.11[b]	−.09[b]	.32[b]	.12	−.14	.38	.11	21	88	.09	−.18	.35	.12	23	86
Frustration	15	.23[b]	.10	.37	.24	.06	.42	.09	25	86	.21	.03	.39	.09	25	84

Note. FE results are replicated results; % Under means the percentage underestimation comparing the FE CI to the RE CI; % CI means the true confidence level (%) comparing the FE CI to the RE CI.

[a] For this mean value, the weights of studies that had *N*s greater than 90 were modified by specifying that $N = 90$ (Bettencourt & Miller, 1996, p. 433). This applies to all subsequent analyses.

[b] The replication of original results is off by .01, which can be attributed to rounding errors.

[c] The *k* in the original was 16.

[d] The values in the original are −.13, −.28, and .03 from the left to the right.

Table 5. Reanalyses of Tables 9, 10, and 11 of Bettencourt and Miller (1996)

		Fixed-effects model			Random-effects models											
		Hedges–Olkin			Hunter–Schmidt						Hedges–Vevea					
Category	*k*	$\hat{\bar{\delta}}$	95% CI		$\hat{\bar{\delta}}$	95% CI		Diff	% Under	% CI	$\hat{\bar{\delta}}$	95% CI		Diff	% Under	% CI
							Table 9									
Neutral condition																
Physical attack	35	.36	.26[a]	.47	.41	.20	.62	.21	50	63	.44	.24	.65	.20	49	57
Verbal aggression	13	.30	.12	.47[a]	.34	.04	.65	.26	43	72	.33	.04	.61	.22	39	76
Provocation condition																
Physical attack	26	.29[a]	.16[a]	.42[a]	.30	.15	.45	.04	13	91	.27	.12	.43	.05	16	89
Verbal aggression	20	.06[a]	−.07[a]	.18	.04	−.16	.24	.15	38	77	.02	−.16	.21	.12	32	78
							Table 10									
							Gender difference effect size									
Neutral condition																
Male target	14	.29	.12	.46	.30	.04	.56	.18	35	80	.27	.00	.54	.20	37	78
Same as participant	19[b]	.32[c]	.18[c]	.47[c]	.39	.06	.73	.38	57	57	.55	.22	.88	.37	56	30
Provocation condition																
Female target	12[b]	.21	.04	.37	.22	−.03	.46	.16	33	81	.19	−.07	.44	.18	35	79
Male target	18	.24	.10	.37[a]	.24	.09	.39	.03	10	92	.23	.07	.38	.04	13	91
Same as participant	17[b]	.07	−.09[a]	.23[a]	.05	−.26	.36	.30	48	68	−.05	−.35	.26	.29	48	57
							Provocation difference effect size									
Female participants																
Female target	23[b]	.79[d]	.64[d]	.94[d]	1.06	.58	1.53	1.10	68	27	1.37	.91	1.83	.62	67	3
Male participants																
Male target	23[b]	.84[e]	.71[e]	.98[e]	.91	.59	1.24	.78	58	55	.74	.38	1.09	.44	62	47
							Table 11									
							Gender difference effect size									
Neutral condition																
Male experimenter	13	.32	.14	.50	.33	.11	.56	.09	20	88	.32	.09	.55	.10	22	87

Table 5. (*Continued*)

		Fixed-effects model			Random-effects models										
		Hedges–Olkin			Hunter–Schmidt					Hedges–Vevea					
Category	*k*	$\hat{\hat{\delta}}$	95% CI		$\hat{\hat{\delta}}$	95% CI		Diff	% Under	% CI	$\hat{\hat{\delta}}$	95% CI		Diff	% Under	% CI
					Provocation difference effect size											
Female participants																
Female experimenter	11	1.28[f]	1.00[f]	1.55[f]	1.70	.86	2.55	1.97	67	31	1.69	.76	2.61	1.30	70	31
Male participants																
Male experimenter	14	.82[g]	.63[g]	1.01[g]	.94	.44	1.43	1.18	62	50	.81	.30	1.32	.64	63	53

Note. FE results are replicated results; % Under means the percentage underestimation comparing the FE CI to the RE CI; % CI means the true confidence level (%) comparing the FE CI to the RE CI.

[a] The replication of original results is off by .01, which can be attributed to rounding errors.

[b] The *k* in the original were 20, 11, 18, 24, and 24 from the top to the bottom.

[c] The values in the original are .38, .24, and .52 from the left to the right.

[d] The values in the original are 1.04, .87, and 1.20 from the left to the right.

[e] The values in the original are .74, .59, and .88 from the left to the right.

[f] The values in the original are 1.24, .98, and 1.50 from the left to the right.

[g] The values in the original are .70, .50, and .90 from the left to the right.

Table 6. From Table 2 of Byrnes *et al.* (1999)

		Fixed-effects model			Random-effects models										
		Hedges–Olkin			Hunter–Schmidt						Hedges–Vevea					
Task type	*k*	$\hat{\delta}$	95% CI		$\hat{\delta}$	95% CI		Diff	% Under	% CI	$\hat{\delta}$	95% CI		Diff	% Under	% CI
Hypothetical choice																
Choice dilemma	44	.07	.02[a]	.12	.07	−.02	.17	.09	47	70	.06	−.03	.16	.09	47	69
Framing	27	.04[b]	−.02	.11[b]	.04	−.05	.14	.05	32	82	.05	−.05	.14	.06	32	82
Other	26[c]	.35	.29	.42[b]	.36	.25	.46	.08	38	77	.37	.25	.48	.10	43	72
Self-reported behaviour																
Smoking	10	−.02	−.04[b]	.01	−.02	−.08	.04	.06	58	58	−.10	−.19	−.01	.13	72	9
Drinking/drug use	18[c]	.04	.02	.06	.04	.00	.08	.04	50	67	.06	.00	.12	.08	67	40
Sexual activities	47	.07	.04[b]	.09[b]	.07	−.03	.16	.14	74	39	.06	−.03	.16	.14	74	39
Driving	21	.29	.26	.32	.29	.20	.39	.13	68	46	.33	.23	.44	.15	71	33
Other	36[c]	.32[d]	.30[d]	.35[d]	.33	.24	.42	.12	72	41	.38	.27	.48	.16	76	22
Observed behaviour																
Physical activity	12[c]	.16	.10	.22	.17	.01	.32	.19	61	55	.30	.13	.48	.23	66	17
Driving	15[c]	.18[b]	.14[a]	.21[b]	.18	.11	.24	.03	46	70	.17	.07	.28	.14	67	48
Informed guessing	12[c]	.18	.13	.23	.18	.02	.35	.23	70	45	.17	−.02	.36	.28	74	39
Gambling	35[c]	.22[b]	.15[b]	.28	.22	.12	.32	.06	35	80	.20	.10	.31	.08	38	76

Note. FE results are replicated results; % Under means the percentage underestimation comparing the FE CI to the RE CI; % CI means the true confidence level (%) comparing the FE CI to the RE CI.

[a] The values in the original are .05 and .12.

[b] The replication of original results is off by .01, which can be attributed to rounding errors.

[c] The *k* in the original were 25, 19, 35, 11, 14, 11, and 33 from the top to the bottom.

[d] The values in the original are .38, .35, and .41 from the left to the right.

Table 7. Reanalyses of Tables 5, 6, and 10 of Bettencourt *et al.* (2006)

		Fixed-effects model			Random-effects models										
		Hedges–Olkin			Hunter–Schmidt						Hedges–Vevea				
Category	*k*	$\hat{\delta}$	95% CI		$\hat{\delta}$	95% CI		Diff	% Under	% CI	$\hat{\delta}$	95% CI		Diff	% Under	% CI
							Table 5									
Trait aggressiveness																
Neutral[a]	12[b]	.94	.78	1.09	1.23	.57	1.89	1.00	76	25	.85	.28	1.43	.83	72	40
Provoking[a]	17[b]	.73	.62	.84	.75	.57	.93	.14	38	76	.73	.53	.93	.17	44	73
Trait irritability																
Provoking[a]	10	.95	.77	1.13	1.18	.52	1.84	.96	73	33	1.16	.55	1.77	.86	70	36
Type A personality																
Neutral[c]	10	−.08	−.29	.14	−.07	−.30	.16	.04	9	93	−.06	−.29	.17	.04	9	92
Provoking[c]	12	.47	.29	.66	.48	.29	.67	.01	2	94	.47	.28	.67	.01	3	94
							Table 6 Physical									
Provocation sensitive																
Neutral[c]	24	.06	−.05	.17	.06	−.07	.20	.04	15	90	.06	−.08	.20	.05	18	89
Provoking[c]	33[b]	.50	.40	.60	.52	.40	.63	.03	13	90	.51	.40	.63	.03	15	90
Aggression prone																
Neutral[a]	17	.87	.74	1.00	1.12	.61	1.64	.77	75	25	.80	.36	1.24	.62	70	42
Provoking[a]	24[b]	.78	.68	.87	.86	.60	1.12	.34	64	44	.84	.60	1.08	.30	61	50
							Table 10 Forced to aggress									
Provocation sensitive																
Neutral[d]	14[b]	.07	−.09	.24	.08	−.11	.26	.04	10	92	.08	−.11	.26	.04	11	92
Provoking[a]	19	.56	.43	.70	.58	.41	.75	.07	19	88	.56	.39	.74	.07	22	88
Aggression prone																
Neutral[a]	14	1.09	.93	1.25	1.40	.79	2.01	.90	74	25	.88	.33	1.43	.78	71	33
Provoking[a]	16[a]	.84	.70	.98	.99	.57	1.42	.58	68	38	.89	.50	1.27	.50	65	50

Table 7. (*Continued*)

		Fixed-effects model			Random-effects models										
		Hedges–Olkin			Hunter–Schmidt					Hedges–Vevea					
Category	*k*	$\hat{\bar{\delta}}$	95% CI		$\hat{\bar{\delta}}$	95% CI		Diff	% Under	% CI	$\hat{\bar{\delta}}$	95% CI		Diff	% Under	% CI
								Free to aggress								
Provocation sensitive																
Neutral[d]	12[b]	.06	−.09	.21	.06	−.13	.25	.08	20	88	.06	−.14	.26	.10	25	86
Provoking[c]	19	.49	.36	.62	.50	.37	.63	.00	0	95	.48	.36	.60	−.01	−6	96
Aggression prone																
Provoking[a]	11	.74	.61	.87	.77	.53	1.01	.22	46	70	.82	.55	1.08	.27	51	59

Note. FE results are replicated results; % Under means the percentage underestimation comparing the FE CI to the RE CI; % CI means the true confidence level (%) comparing the FE CI to the RE CI.

[a] The values in the original are different from our replication results due to difference in *k*s and/or their non-traditional use of sample size (rather than inverse sampling error variance) as weight for the FE methods (p. 761).

[b] The *k* in the original were 11, 16, 34, 23, 15, 25, and 11 from the top to the bottom.

[c] The replication of original results is perfect or off by .01, which can be attributed to rounding errors.

[d] The replication of original results is off by .02 or less.

Table 2 presents the results for the Hyde and Linn (1988) study. The FE model underestimates the CI width in 9 of the 10 meta-analyses. The average percentage underestimation is 55% according to the HS RE procedure and 65% according to the HV RE procedure. Judged against the HS RE CIs, the nominal 95% FE CIs were on average 57% CIs. Results were more discrepant for the HV RE procedure, which indicates that the nominal 95% FE CIs were really on average only 33% CIs. Both comparisons, but especially that with the HV procedure, indicate serious underestimation of CIs by the FE model.

Table 3 shows the reanalysis for the Hyde *et al.* (1990) meta-analysis. The general pattern is the same as in the previous reanalysis. For the HS RE procedure, the average percentage by which FE CIs underestimated actual CIs is 70%, which is again a serious inaccuracy. The nominal 95% FE CIs are on average actually 40% CIs. The results for the HV RE model are again more extreme: an average of 77% underestimation of the CIs and an average 20% confidence level on average. Again, the two RE procedures agree in indicating that the FE CIs are seriously in error.

The data from the Bettencourt and Miller (1996) meta-analysis were extensive enough to require two tables. Results from this study are presented in Tables 4 and 5. In Table 4, results for the HS RE procedure indicate that the FE CIs underestimate the actual CI widths on average by 41%. For the HV RE procedure, the average underestimate is 39%. In comparison to the HS RE CIs, the nominal 95% FE CIs are on average 67% CIs; the corresponding figure for the HV RE procedure is 68%. In these data, the results given by the two RE models are nearly identical. And again, both RE procedures indicate that the FE CIs are seriously in error: CI width is too narrow by about 40% and nominal 95% CIs are really on average only 68% CIs.

The results shown in Table 5 for the remainder of the data from Bettencourt and Miller (1996) are similar to the Table 4 results. Again, the results for the two RE models are almost identical. Results for the HS RE procedure indicate that on average the FE CIs underestimate the actual CI width by 43%; the corresponding figure for the HV RE procedure is 44%. In comparison to the HS RE CIs, the nominal 95% FE CIs are on average only 67% CIs; for the HV RE procedure, this figure is 61%. Again, the key fact is that two different RE models both indicate that the FE CIs are quite inaccurate.

Table 6 presents the reanalysis for the Byrnes *et al.* (1999) meta-analysis. The overall pattern of results is again very similar. Based on the HS RE model, the FE CIs underestimated the real CIs by 54%. The nominal 95% FE CIs were on average only 61% CIs. Results for the HV RE model were a little more extreme: 61% underestimation of the CIs on average, and a 46% CI on average. Again, the two RE procedures agree in indicating that the FE CIs are much narrower than the actual CIs and do not come close to attaining the 95% coverage that a 95% CI should have.

The results for the 16 meta-analyses from Bettencourt *et al.* (2006) study are presented in Table 7. The FE model underestimates the CI width for 15 of the 16 meta-analyses. The average percentage underestimation is 38% according to both the HS and HV RE procedures. Evaluated against the HS RE CIs, the nominal 95% CIs were on the average 67% CIs. For the HV RE procedure, this figure was 69%. In this set of data, the two RE procedures yielded very similar results. The discrepancies between the FE procedure and the two RE procedures were again substantial, although less extreme than in some of the previous tables.

To provide a summary picture of the results, we averaged the results across Tables 2–7. In comparison to the HS RE CIs, the FE CIs underestimated the width of the actual CIs by 50% on average; for the HV RE procedure, this figure was 53%. Hence, on average the two RE procedures produce similar verdicts on the FE CIs. The FE CIs are on average

less than half as wide as the actual CIs. The average underestimation across the two RE models is 52% (rounded). This is obviously a large discrepancy.

We also averaged the confidence levels of the FE CIs across Tables 2–7. Using the HS RE procedure as the standard, the nominal 95% FE CIs are on average only 60% CIs. Using the HV RE procedure as the standard, the nominal FE CIs are really 51% CIs on average. The difference here between the two RE procedures seems larger than for the percentage underestimation figures, but the greater discrepancy is predicted by the properties of the normal curve that is the basis for the CIs. That is, small differences in CI width between the two RE procedures result in larger differences in percentage coverage (confidence levels) because of normal curve properties. On average across the two RE procedures, the nominal 95% FE CIs are estimated to be 56% CIs. Hence, on average FE CI coverage is only about 59% of its nominal value (i.e. .56/.95 = .59). Again, this is a major discrepancy.

While estimates of $\bar{d}$ were typically quite similar for the HS and HV RE models, they were sometimes substantially different. For example, in Table 2 for reading comprehension, the means were .03 and .09, respectively. And in Table 3 for problem solving, the means were .09 and .19, respectively. Other examples are apparent in the tables. These differences are due to the difference in the study weights used to compute the means. The HS model weights each study by its sample size (N_i), while in the HV model, the study weights are $w_i^* = 1/(V_{e_i} + \hat{\sigma}_\delta^2)$. The larger $\hat{\sigma}_\delta^2$ is, the more different these two sets of study weights are. As $\hat{\sigma}_\delta^2$ becomes larger, the HV weights become less unequal across studies, while this does not happen for the HS weights.

In addition to the value of $\hat{\sigma}_\delta^2$, a correlation between N_i and the d_i statistic, $r(N_i, d_i)$, could cause the two weighting approaches to produce different estimates of the mean. In our five meta-analyses, the average correlation between N_i and d_i ranged from $-.10$ to $+.10$, with a grand mean of $-.01$. This would suggest random variation around a mean of zero. This hypothesis was confirmed by the finding that the variation across studies in these correlations was on average no larger than expected on the basis of sampling error alone. Using the HS approach, we found that on average across the five meta-analyses all variation in these *r*s was attributable to sampling error variance. Using the HV approach, we found that all the homogeneity tests (*Q* tests) except one were non-significant.

A frequently used measure of publication bias is a negative correlation between d_i and N_i, assumed to result from failure to publish small *N* studies with small (and therefore non-significant) *d* values – i.e. a 'file drawer problem' (Rothstein, Sutton, & Borenstein, 2005). The mean correlation of zero suggests the absence of publication bias in these data.

Although this should not be the case (see later discussion), most authors interpret CIs as significance tests. If the CI does not include zero, $\bar{r}$ or $\bar{d}$ is declared statistically significant. Hence, it is clear that erroneously narrow CIs lead to an inflated Type I error rate. Field (2003) presents a computer simulation study that suggests that the Type I error rate is substantially inflated when the FE model is used in meta-analysis. This Type I error problem is discussed from an analytic perspective in Hedges and Vevea (1998), Hunter and Schmidt (2000), and Overton (1998). However, for the reasons given earlier it is not our major focus in this paper.

7. Discussion

This study is the first to use empirical data from the archival literature to compare the results produced by fixed- and random-effects meta-analysis models. It is clear that results differ substantially depending on whether the FE or RE model is used. Results using the RE model indicate that meta-analysis findings are much less exact and precise

than is indicated by the commonly used FE model. In comparison to these large differences, the minor differences between the two RE procedures seem unimportant and in any event are (usually) in the predicted direction. Also, the similarity of results for the two RE models suggests that the positive bias in the HV RE procedure estimate of sampling error variance, but not present in the HS RE procedure, is of limited importance when viewed against the background of the far larger differences between both RE procedures and the FE procedure.

As shown earlier, most meta-analysis studies that have appeared in *Psychological Bulletin* have been based on the FE model (see Table 1 and Figure 1). If we accept the proposition that the goal of research is generalizable knowledge (and not merely knowledge about the specific set of studies in the meta-analysis) and if we accept the National Research Council's (1992) interpretation of FE and RE models, we are led to conclude that most of the meta-analysis results in the leading US psychology review journal may be substantially in error in their statements of precision of findings. Although this paper does not explore this question in any detail, we are also led to conclude that Type I errors may be quite frequent in the meta-analysis literature in some research areas when researchers use FE methods and interpret CIs as significant tests. The problem may seem potentially less serious if we accept the proposition by Hedges and Vevea (1998) that, by analogy with generalized interpretations sometimes made of FE ANOVA experimental data, there is an extra-statistical basis for generalizing FE meta-analysis findings beyond the specific studies in the meta-analysis. However, as Hedges and Vevea (1998) indicate, adoption of this notion requires an ascertainment, based on an extra-statistical subjective judgement, that studies not included in the meta-analysis are 'sufficiently similar' to those included to justify generalization. However, we found that the question of such similarity was not explicitly addressed in any of the 68 meta-analyses we reanalysed nor in any of the 129 FE meta-analysis studies in *Psychological Bulletin*; and it is not clear how it would or should be approached (Schulze, 2004). In any event, we are still left with the difficulty that the National Research Council report, written by a select group of statisticians appointed by the National Science Foundation, has rejected this analogy and this interpretation of FE models. Even if this analogy were accepted, the question would still remain of whether the broad generalization of findings of FE ANOVA-based experiments that are sometimes made by primary researchers (Hedges & Vevea, 1998) is justified. As noted earlier, the fact that they are sometimes made does not *per se* constitute a justification for making them (Schulze, 2004).

7.1. Implications for research, practice, and policy

The present findings have potentially important implications for researchers, practitioners, and policy makers. The CI is often used in statistical inference, with the decision being that an effect is real if the CI does not include zero. The narrower CIs of the FE model are more likely to exclude zero when the more accurate CIs of the RE model would include zero. For example, in Table 2 the RE CI for reading comprehension includes zero while the FE CI does not. Thus, given common approaches to data interpretation, the FE model leads to the conclusion that females have better reading comprehension than males while the RE model does not. This sort of difference can occur in any area of research. In light of what has been presented in this paper, it is likely that the RE-based interpretation is correct and the FE-based interpretation is not. Furthermore, even in cases in which neither the FE nor the RE CI includes zero, the level of uncertainty about the mean value can play an important role in practical decisions. If the CI (referred to in lay terms as the 'error band' and familiar to the public from its use

in opinion polls) is narrow, as it is more likely to be with the FE model, researchers and policy makers may be overly confident that they have accurate and 'hard' information to act on. On the other hand, if these same users were exposed to the more accurate and often much wider RE CIs they may rightfully be considerably more cautious in their decision making. In fact, one reason why primary researchers have been reluctant to substitute CIs for significance tests is probably that CIs are often wide and hence reveal just how little information the individual primary study contains (Hunter & Schmidt, 2004). This consideration will likely become more important in the future as the movement to educate researchers and others in the proper interpretation of CIs becomes increasingly successful (American Psychological Association, 2001; Belia, Fidler, Williams, & Cumming, 2005; Schmidt, 1996; Thompson, 2002, 2006). The goal of this movement is to wean researchers and policy makers from naïve dichotomous thinking that looks only to see whether an effect or relation is statistically significant or not to a focus on the *magnitude* of the estimate of the effect and the *precision* of that estimate of magnitude. To the extent that this effort to reform data analysis and interpretation procedures is successful, there will be a greatly increased emphasis in the future on the width of CIs in data interpretation. The result will be increased importance for accurate estimates of CIs.

7.2. Generalization of findings to the correlation coefficient

The findings and conclusion of this paper can safely be generalized to meta-analyses in which the summary statistic is the (Pearson) correlation coefficient instead of the *d*-value statistic. That is, as with the *d*-value statistic, the FE model will result in CIs that are too narrow when evaluated against the more accurate RE CIs. Except for the fact that the simple sampling error variance formula for the correlation coefficient is different, the RE procedures for the correlation are identical to those for the *d*-value statistic. Therefore, the cautions we express against the routine use of the FE model with *d* values also apply to correlations. As discussed earlier, the only reason why we did not use *r*-based meta-analyses to compare the FE model to the two RE procedures is that the HV procedures (both RE and FE) conduct the analysis using the Fisher *z* transformation of *r*, while the HS procedure does not. This transformation has little effect on the accuracy of mean estimates in the FE model but leads to upward biases in estimates of mean *r* in the RE model. Hence, in most RE applications, the HV RE model will indicate larger mean *r* estimates than the HS RE procedure. In the present study we thought it wise to avoid the distraction that would be created by this difference and so chose to focus on studies using the *d*-value statistic, which is identical in both RE procedures. Of course, one could apply the HV RE procedure for *r*s without the Fisher *z* transformation (or vice versa), but then one would be departing from one of the procedures as presented by its originators and the analysis might be challenged on that basis. However, it can be confidently stated that an application of the two RE models as presented by their authors would lead to conclusions about CI widths very similar to the present conclusions, because the statistical and mathematical principles are the same. However, the larger differences in the mean *r* estimates might distract attention from this key point.

7.3. Some important technical issues

The question can be raised at this point as to why the reported FE CIs are apparently substantially too narrow if the χ^2 test of homogeneity (the *Q* test; Hedges & Olkin, Ch. 9;

Hedges, 1992) has been used appropriately in FE meta-analyses. Hedges and Olkin (1985) stated that the *Q* test should precede the use of the FE model. If this test is non-significant, the hypothesis of homogeneity of study population parameters is accepted and use of the FE model can be supported (implying that FE and RE procedures would produce the same results). If this test is significant, the conclusion is that the variance of study effect sizes is larger than can be explained by simple sampling error and therefore the study population values of ρ or δ are deemed variable, indicating the presence of moderator variables or interactions. In such cases, Hedges and Olkin (1985) suggested that the studies should be subdivided based on potential moderators into subsets that have within-set homogeneous study population parameters, as indicated by non-significant *Q* tests. However, a non-significant homogeneity test, whether before or after subsetting the studies, does not provide reliable support for a conclusion of homogeneity. Unless the number of studies is large, this χ^2 test typically has low statistical power to detect variation in study population parameters, resulting in frequent Type II errors (Hedges & Pigott, 2001; Mengersen, Tweedie, & Biggerstaff, 1995; Morris & DeShon, 2002; National Research Council, 1992, p. 52; Schulze, 2004, p. 195). That is, the χ^2 is often non-significant in the presence of real variation in study population parameters (Hedges & Olkin, 1985). As a result, FE models may be applied to heterogeneous sets of studies, resulting in CIs that are substantially too narrow.

In addition, even if the *Q* test is significant (indicating heterogeneity of study population values), published meta-analysis studies often nevertheless apply the FE model, making it even more likely that the resulting CIs will be too narrow. We identified 38 meta-analysis studies published in *Psychological Bulletin* between 1980 and January 2006 that followed this practice. This is 29% of the 129 meta-analysis studies that used FE methods. Four of the five meta-analysis studies which we reanalysed followed this practice (one did not apply the *Q* statistic at all.) The 29% figure is an underestimate, because in many studies authors responded to initially significant *Q* statistics by subgrouping studies by potential moderators and computing new *Q* statistics. When these were again significant, there was no further subgrouping of studies, and the FE model was then used and interpreted despite the significant *Q* statistics. We did not include these studies in our count of 38.

The percentage underestimation of the CI by the FE model should be greater when the homogeneity test (*Q*) is statistically significant than when it is not. The *Q* test was non-significant in 24% of the 68 meta-analyses. The correlation between the significant–non-significant dichotomy and percentage underestimation of the CI was .75 for the HS RE model and .78 for the HV RE model. We can also look at the relation between $\hat{\sigma}_\delta^2$ and percentage underestimation of the CI. The average correlation across the five meta-analysis studies between $\hat{\sigma}_\delta^2$ and percentage underestimation of the CI width by the FE model is .66 for the HS RE model and .61 for the HV RE model. For the square root of $\hat{\sigma}_\delta^2$ (i.e. the estimated *SD* of the population parameters), these correlations are somewhat higher, as would be expected: .71 and .65, respectively. These relationships are in the expected direction and are substantial.

When between-study variance in population parameters is large, the value of presenting the estimated mean effect size and the CI for the mean can be questioned, at least for most theory-testing purposes. Hunter and Schmidt (2004) state that, for this reason, in such cases their full procedure recommends presentation of *credibility* intervals (CrIs), not *confidence* intervals (CIs). The CrI refers not to the mean (as the CI does) but to the estimated distribution of the study population parameters. For example, the 90% CrI includes the middle 90% of values in the estimated distribution of

population values. Of course, it is important to search for moderators (interactions) when the estimated between-study variance in population parameters is large. However, searching for moderators does not guarantee an important reduction in the variance of population parameters. In many cases in the literature, tests of potential moderators do not lead to much reduction in between-study variance, even when the moderator means are at least somewhat different. So, one is then left with the question of whether the CI around the mean value provides much useful information. In such a case, the RE CI will still be more accurate than the FE CI and will better reveal the true uncertainty in the estimate of the mean. However, because the mean is limited in its ability to describe the distribution, a CrI may be more informative in such cases. Nevertheless, the most common procedure followed in meta-analyses in *Psychological Bulletin* (and probably most other journals) is to present mean estimates and CIs around these means (or significance tests of these means). Given these practices, the present paper demonstrates empirically that when the FE model is used (by far the majority of cases; see Table 1) the resulting CIs are substantially too narrow.

7.4. Choice of a model of meta-analysis

Are there any circumstances in which the choice of the FE model would be appropriate? These circumstances would appear to be very limited. The FE model would be appropriate if one had strong evidence that the primary studies to be included in the meta-analysis were virtually identical, i.e. they are all literal or operational replications of each other (Aronson, Ellsworth, Carlsmith, & Gonzales, 1990). That is, if the studies drew their samples from the same population (e.g. college sophomores), tested exactly the same hypotheses with exactly the same study design, treatment strength (if an experimental study), measures, instructions, time limits, etc, then one might assume *a priori* that the same population parameter was estimated in all the primary studies (i.e. σ_δ^2 or $\sigma_\rho^2 = 0$) and this could be a basis for choosing the FE model. Such a situation would be expected to occur only rarely (Aronson *et al.*, 1990). In any other situation, an FE model would be inappropriate and the recommendation would be that any meta-analysis conducted using the FE model should be reanalysed using an RE model. As noted earlier, Overton has presented another rationale for choice of the FE model. He argues that if one has reason to believe that S_δ^2 or S_ρ^2 is small (near zero or zero), choice of the FE model might be justified as a way of avoiding the upward bias in the *SE* estimate (and hence the CI estimate) in the HV and related RE methods. The empirical results in this study suggest that this bias is not large, and in any event does not occur in the HS RE procedure (which was not examined by Overton, 1998). Hence the situations in which choice of the FE model is defensible seem limited.

We can use the present database to provide a very tentative and preliminary estimate of the frequency with which the FE model would be appropriate in the meta-analysis literature. The FE model is appropriate whenever study population parameters have zero or near-zero variance across studies. From an operational point of view, such situations can be identified in this paper as those in which the FE and RE CIs are equal in width. In the 68 meta-analyses represented in Tables 2–7, this occurs only twice, for a frequency of 3%. That is, if we take these data as representative (and they may not be), then they suggest that the FE meta-analysis model is appropriate in only 3% of meta-analyses and inappropriate in 97%. This estimate, combined with the preponderance of the FE model in US psychology's premier review journal, suggests there is a widespread misconception that the FE model is appropriate when it is not. In this sense, the present study is akin to

recent studies demonstrating the existence of widespread misconceptions with respect to other statistical issues. For example, Belia *et al.* (2005) showed that misconceptions regarding the interpretation of confidence intervals are common among researchers, and Oaks (1986; see also Schmidt, 1996) showed this to be the case with respect to statistical significance tests. Improvements in data analysis practices in psychological research require that all such misconceptions be addressed and corrected.

8. Limitations of this study

The major limitation of this study stems from the relatively small sample of FE meta-analyses that it was possible to reanalyse using RE models. It is probably best to view this reanalysis itself as an application of RE model; that is, the meta-analyses we reanalysed can perhaps be viewed tentatively as a sample of all such meta-analyses that could theoretically be reanalysed. As such, our sample of five studies could be unrepresentative. However, our reanalysis included multiple meta-analyses from each of the five published meta-analysis studies, widening the sample to 68 meta-analyses with $k = 10$ or more. On the other hand, these 68 meta-analyses cannot be assumed to be fully independent. However, in light of the clear differences in the statistical properties of FE and RE models, as presented earlier in this paper and elsewhere, and in light of the rarity of the empirical research conditions under which the FE model is appropriate, we believe it is unlikely that conclusions would be materially different with a different or larger sample of FE meta-analyses. In this connection, the key question is probably whether the meta-analyses examined are typical or representative *methodologically*, not whether they are representative in terms of subject matter or area of research. In this connection, the meta-analysis included in our study that was from a substantive area other than gender differences yielded results similar to those from some of the meta-analyses from the area of gender differences. Based on our examination of the many meta-analyses published in *Psychological Bulletin* in connection with the present research, we judge the meta-analyses we examined to be quite typical methodologically. That is, they applied the standard Hedges and Olkin (1985) FE method and did so in the typical manner. They seem to be (usefully) atypical only in that they presented all data needed to replicate their meta-analyses – something we found, to our disappointment, to be quite rare.

9. Conclusion

Meta-analysis is the major tool today in psychology, the social sciences, medicine, and other areas (Hunter & Schmidt, 2004) for revealing the cumulative knowledge contained in research literatures. It has revolutionized the basis for the production of knowledge through research. Yet even today, 30 years after its introduction, important technical issues remain in meta-analysis methods. This study sheds light on what appears to be not only an important technical problem but also an important epistemological problem in the psychological literature: the precision and certainty of meta-analysis findings may have been systematically overstated in much of the research literature. Solving this problem will probably not be easy, but it is important that it be addressed. Our recommendation is that future meta-analyses use RE models and that the older FE meta-analyses be reanalysed using RE models to provide accurate results and conclusions.

References

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

Aronson, E., Ellsworth, P., Carlsmith, J., & Gonzales, M. (1990). *Methods of research in social psychology* (2nd ed.). New York: McGraw-Hill.

Bechtel, W. (1988). *Philosophy of science: An overview for cognitive science*. Hillsdale, NJ: Erlbaum.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357–382). New York: Russell Sage Foundation.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*, 389–396.

Bettencourt, B. A., & Miller, N. (1996). Gender differences in aggression as a function of provocation: A meta-analysis. *Psychological Bulletin*, *119*, 422–447.

Bettencourt, B. A., Talley, A., Benjamin, A. J., & Valentine, J. (2006). Personality and aggressive behavior under provoking and neutral conditions: A meta-analytic review. *Psychological Bulletin*, *132*, 751–777.

Brannick, M. (2006). *Comparison of sample size and inverse variance weights for the effect size r*. Paper presented at the 1st annual meeting of the Society for Research Synthesis Methodology, Cambridge, UK.

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*, 367–383.

Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for validity generalization. *Journal of Applied Psychology*, *65*, 543–558.

Cooper, H. (1997). Some finer points in meta-analysis. In M. Hunt (Ed.), *How science takes stock: The story of meta-analysis* (pp. 169–181). New York: Russell Sage Foundation.

Donnor, A., & Rosner, B. (1980). On inferences concerning a common correlation coefficient. *Applied Statistics*, *29*, 69–76.

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

Field, A. P. (2003). The problem in using fixed-effects models of meta-analysis on real world data. *Understanding Statistics*, *2*, 77–96.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*, 444–467.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Hall, S. M., & Brannick, M. T. (2002). Comparison of two random effects methods of meta-analysis. *Journal of Applied Psychology*, *87*, 377–389.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*, 388–395.

Hedges, L. V. (1988). The meta-analysis of test validity studies: Some new approaches. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 191–212). Hillsdale, NJ: Erlbaum.

Hedges, L. V. (1989). An unbiased correction for sampling error in validity generalization studies. *Journal of Applied Psychology*, *74*, 469–477.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics*, *17*, 279–296.

Hedges, L. V. (1994). Statistical considerations. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 29–38). New York: Russell Sage.

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*, 203–217.

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

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society, Series B*, *15*, 193–232.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Beverly Hills, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law*, *2*, 324–347.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, *8*, 275–292.

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

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, University of Iowa.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*, 139–155.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*, 53–69.

Mengersen, K. L., Tweedie, R. L., & Biggerstaff, B. (1995). The impact of method choice on meta-analysis. *Australian Journal of Statistics*, *37*, 19–44.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105–125.

Myers, D. G. (1991). Union is strength: A consumer's view of meta-analysis. *Personality and Social Psychology Bulletin*, *17*, 265–266.

National Research Council (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy of Sciences Press.

Oakes, M. L. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Osburn, H. G., & Callender, J. C. (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*, 354–379.

Phillips, D. C. (1987). *Philosophy, science, and social inquiry*. Oxford: Pergamon Press.

Raju, N. S., & Burke, M. J. (1983). Two new procedures for studying validity generalization. *Journal of Applied Psychology*, *68*, 382–395.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis*. New York: Russell Sage Foundation.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*, 75–98.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Erlbaum.

Rosenthal, R., & Rubin, D. B. (1982a). Further meta-analytic procedures for assessing cognitive gender differences. *Journal of Educational Psychology*, *74*, 708–712.

Rosenthal, R., & Rubin, D. B. (1982b). Comparing effect sizes of independent studies. *Psychological Bulletin*, *92*, 500–504.

Rothstein, H. F., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustment*. Chichester: Wiley.

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Association*, *75*(372), 801–827.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, *6*, 337–400.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529–540.

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 *z* transformation. *Journal of Applied Psychology*, *73*, 665–672.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology*, *78*, 3–13.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Toronto: Hogrefe & Huber.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 24–31.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford.

Toulmin, S. S. (1961). *Foresight and understanding: An enquiry into the aims of science*. New York: Harper.

Received 15 May 2007; revised version received 13 October 2007