Behavior Research Methods, Instruments, & Computers
2003, 35 (4), 504-511

Estimating the mean effect size in meta-analysis: Bias, precision, and mean squared error of different weighting methods

WIM VAN DEN NOORTGATE and PATRICK ONGHENA
Katholieke Universiteit Leuven, Leuven, Belgium

Although use of the standardized mean difference in meta-analysis is appealing for several reasons, there are some drawbacks. In this article, we focus on the following problem: that a precision-weighted mean of the observed effect sizes results in a biased estimate of the mean standardized mean difference. This bias is due to the fact that the weight given to an observed effect size depends on this observed effect size. In order to eliminate the bias, Hedges and Olkin (1985) proposed using the mean effect size estimate to calculate the weights. In the article, we propose a third alternative for calculating the weights: using empirical Bayes estimates of the effect sizes. In a simulation study, these three approaches are compared. The mean squared error (*MSE*) is used as the criterion by which to evaluate the resulting estimates of the mean effect size. For a meta-analytic dataset with a small number of studies, the *MSE* is usually smallest when the ordinary procedure is used, whereas for a moderate or large number of studies, the procedures yielding the best results are the empirical Bayes procedure and the procedure of Hedges and Olkin, respectively.

The standardized mean difference was first used by Cohen (1969) to perform power calculations for *t* tests. Ever since, this measure of effect size has frequently been used to summarize the results of group comparison studies. Glass and his colleagues (Glass, 1976, 1977; Glass, MacGaw, & Smith, 1981), who were among the first to present a systematic set of meta-analytic techniques, suggested using this measure in meta-analysis. The sampling distribution of the effect size measure and the statistical background for its use in meta-analysis was provided by Hedges (1981, 1982). Although other measures of effect size have been advocated in meta-analytic research—for instance, the correlation coefficient (e.g., Hunter, Schmidt, & Jackson, 1982) or Fisher's *Z* (e.g., Rosenthal, 1991)—the standardized mean difference still plays a prominent role in meta-analytic research.

In this article, we discuss a problem that shows up when the overall or mean standardized mean difference is estimated with standard methods. Traditionally, the grand mean effect is usually estimated by a weighted average of the observed effects. The precision of an observed effect size, which is the inverse of the sampling variance, is often used to weight an observed effect size. Owing to the dependence of the weights and the (observed) standardized mean difference, the estimate of the mean standardized mean difference will be biased. As we shall see below, several alternative estimation procedures are possible, but from the literature it is not clear whether there is a uniformly best approach or in which situation one or another approach is to be preferred. One appealing approach has been proposed by Hedges and Olkin (1985). They proposed making the weights independent of the observed effect sizes by using the overall effect size estimate to estimate the weights. Recent developments in meta-analytic literature, however, suggest that empirical Bayes estimates of the study effect sizes are optimal estimates of the study effect sizes, reducing the mean squared error (*MSE*; Raudenbush & Bryk, 1985). Hence, a plausible alternative approach for estimating the sampling variances of the observed effect sizes and the corresponding weights, which however has not been proposed before, is to use the empirical Bayes estimates of the study effect sizes. The purpose of this article is (1) to explore the extent of the problem of bias, (2) to present the empirical Bayes approach for estimating the weights, and (3) to evaluate the mean effect size estimates based on the ordinary procedure, the empirical Bayes procedure, and the overall effect size procedure for estimating the weights. We will start with a short description of the standardized mean difference and of common meta-analytic techniques. Next, we will discuss the problem of estimating the grand mean effect size when standard methods are used and will present some potential improvements of the estimation procedure. Three alternatives will be compared, using a small simulation study. We will finish with some conclusions and recommendations.

The Standardized Mean Difference

The standardized mean difference, δ, is defined as the difference between two population means divided by the standard deviation of one of both populations or by a common population standard deviation (σ). It is often used

Correspondence concerning this article should be addressed to W. Van den Noortgate, Department of Education, Katholieke Universiteit Leuven, Vesaliusstraat 2, B-3000 Leuven, Belgium (e-mail: wim.vandennoortgate@ped.kuleuven.ac.be).

when an experimental population is compared with a control population, in order to investigate the effect of a certain intervention:

$$\delta = \frac{\mu_E - \mu_C}{\sigma}, \quad (1)$$

where μ_E and μ_C the means of the experimental and the control populations, respectively.

This measure is sometimes called Cohen's d, but because it is a population parameter, we prefer using the Greek symbol δ. Hedges and Olkin (1985) argued that in experimental research, it is often sensible to assume a common population variance. An unbiased estimate of this variance can be obtained by pooling the two sample estimates:

$$s_p^2 = \frac{(n_E - 1)s_E^2 + (n_C - 1)s_C^2}{N - 2}, \quad (2)$$

where n_E is the size of the experimental group, n_C is the size of the control group, and N is equal to $n_E + n_C$.

An obvious estimate of δ, therefore, is its sample counterpart g:

$$g = \frac{X_E - X_C}{s_p}. \quad (3)$$

Although the sample mean difference is an unbiased estimator of the population mean difference and the pooled variance estimate is an unbiased estimate of a common population variance, g is a biased estimator of δ, especially for smaller studies (Hedges, 1981). An unbiased estimator d can be obtained by multiplying g by a correction factor. Hedges (1981) found that d is very closely approximated by

$$d \cong g \cdot \left[1 - 3 / (4N - 9)\right]. \quad (4)$$

Hedges (1981) showed that if both the experimental and the control population distributions are normal and n_E and n_C are moderate to large (at least equal to 10), the sampling distribution of d is approximately normal:

$$d \sim N\left[\delta, \sigma^2(d)\right],$$

where

$$\sigma^2(d) \cong \frac{N}{n_E n_C} + \frac{\delta^2}{2N}. \quad (5)$$

The sampling variance depends on the unknown population parameter δ but is closely approximated by replacing δ in Equation 5 by the estimate of δ:

$$d \sim N\left[\delta, \hat{\sigma}^2(d)\right],$$

where

$$\hat{\sigma}^2(d) = \frac{N}{n_E n_C} + \frac{d^2}{2N}. \quad (6)$$

Combining Effect Size Estimates From Different Studies

In meta-analytic research, it is often assumed that, in each study, the same effect size δ is estimated. This δ is estimated by averaging the observed effect sizes. Because studies often differ in size and, therefore, the precision of the estimates differs, usually a weighted mean of the observed effect sizes is calculated. By using weights equal to the precision (the inverse of the variance calculated using Equation 6), the variance of $\hat{\delta}$ is reduced (e.g., Hedges, 1982):

$$\hat{\delta} = d_+ = \frac{\sum_{i=1}^{k} w_j d_j}{\sum_{j=1}^{k} w_j}, \quad (7)$$

where $w_j = 1/\hat{\sigma}^2(d_j)$ and $k =$ the number of studies.

Often, however, the assumption of homogeneous studies (i.e., studies in which a common effect size is estimated) is questionable, because studies usually are not exact replications of each other. Although in this case the fixed effects procedure presented above is still appropriate if the true effects investigated in the set of studies are regarded as the effects of interests (Hedges & Vevea, 1998), a random effects meta-analytic model may be more appropriate. In a random effects meta-analytic model, the true effect sizes in a set of studies are viewed as a random sample out of a population of effect sizes (Rubin, 1981). Consequently, in this model, there are two sources of variation of observed effect sizes: variation between true effect sizes and sampling variation for each study. Because these two sources are independent,

$$\sigma^2(d_j) = \sigma^2(\delta) + \sigma^2(d_j \mid \delta_j). \quad (8)$$

Meta-analytic procedures for random effects models (REMs) were further developed by Hedges (1983; Hedges & Olkin, 1985) and DerSimonian and Laird (1986). The mean true effect size is also estimated using a precision-weighted average of the observed effect sizes (Equation 7), but the variance of the observed effect sizes (and therefore, the weights) is now calculated on the basis of Equation 8. Raudenbush and Bryk (1985) showed that a meta-analysis is a special case of a multilevel analysis and that estimation procedures for multilevel models (usually maximum likelihood procedures) can be used to estimate the unknown parameters of the meta-analytic models. The different approaches may lead to different between-studies variance estimates, especially if the number of studies is small, and hence to slightly different estimates of the mean effect size and to different estimates of the standard errors of estimation. Van den Noortgate and Onghena (2003) showed that the DerSimonian and Laird and the maximum likelihood estimates are usually very similar and are, in general, to be preferred over the between-studies variance estimates proposed by Hedges and Olkin.

In both the fixed effects model (FEM) and the REM, study characteristics can be included as moderator variables, but the discussion in this article will be restricted to models without moderator variables.

The Use of the Standardized Mean Difference in Meta-Analysis

The use of the standardized mean difference in meta-analysis is appealing for different reasons, including the

ease of interpretation: d is equal to the difference between the means, after correction for different scaling. Indeed, if in different studies different but linearly equatable scales are used, dividing by the standard deviation makes the mean differences directly comparable. Moreover, d is well known among many researchers, and formulas are widely available in the meta-analytic literature to convert a large range of effect size measures, test statistics, or p values to d. Nevertheless, the standardized mean difference also has some drawbacks. One of them is its restricted applicability: d can be used to compare two groups or to describe the effect of a dichotomous variable but cannot be used for continuous independent variables, in contrast with, for instance, the correlation coefficient.

Another problem, which has often been overlooked by meta-analysts, is the main focus of this article. The problem was pointed out by Hedges (1983; Hedges & Olkin, 1985) in cases of homogeneous studies. As was discussed above, under the assumption of normally distributed populations with a common variance, d gives an unbiased estimate of δ for each study. Consequently, the same is true for the unweighted mean of the observed effect sizes stemming from a random set of homogeneous studies. However, the weighted combination of ds can be a biased estimate of δ. This is the case in the common procedure, in which the inverses of the sampling variances are used as weights and the sampling variances are estimated by replacing δ with d (Equation 6). This procedure implies that for large positive or negative ds, the estimated sampling variance will be somewhat larger than that for small ds, resulting in a smaller weight. This means that, in general, the estimate of a nonzero δ based on a set of observed ds is shrunken toward zero, resulting in a negative bias for estimating a positive δ and a positive bias for a negative δ. Because the weights are less dependent on δ if the study sizes are large and because the variation of the observed effect sizes is smaller for a small true effect size (see Equation 6), we would expect that the problem is largest for small studies or studies estimating a large effect. In contrast, because the weights do not depend on the number of studies, we would expect that the bias is independent of the number of studies. In cases of heterogeneous studies, we would expect that the bias is even larger, because in this case the variance of the effect sizes is due not only to sampling variation, but also to the variance between true effect sizes. This means that in addition to the bias that is due to the fact that the sampling variance is not known but is estimated using the observed effect size, the estimate is biased because of the true between-studies variance. Finally, it can be deduced from Equations 7 and 8 that using an REM, instead of an FEM, can partially compensate for the bias, especially when the true variance is large: Because in the random effects procedure, not only the sampling variance, but also the true between-studies variance is taken into account in calculating the weights, the weights are less dependent on the sampling variances and are, thus, more equalized. In the next section, possible solutions for the problem of bias will be discussed.

Improving Mean Delta Estimates

As Hedges (1983) has pointed out, using weights that are independent of δ (e.g., $n_E n_C/(n_E + n_C)$ solves the problem of the bias in estimating a common δ, but the resulting estimator is less efficient. Another, not frequently used, alternative is to employ a fixed estimate of δ for estimating the sampling variance in all studies. This can be an a priori estimate of δ or an estimate resulting from a meta-analysis. Because the estimates of the sampling variances are needed for a meta-analysis, Hedges and Olkin (1985) proposed an iterative meta-analysis. A meta-analysis is performed with weights estimated using the observed effect sizes, and the resulting estimate of the mean true effect size is used to estimate the weights for a second meta-analysis, and so on. Weights thus are updated in each successive meta-analysis. Note that Hunter and Schmidt (1990) proposed a similar (but noniterative) approach for combining correlation coefficients: The observed effect sizes are weighted with their corresponding reliabilities, which are estimated using the average correlation across studies.

The iterative procedure outlined by Hedges and Olkin (1985) may raise problems if the meta-analytic dataset contains heterogeneous studies, because in this case, the weighted mean of the effect sizes (d_+) gives an estimate of the mean effect size that may be relatively uninformative for the true effect sizes of the individual studies. In case studies that are strongly heterogeneous, estimating the sampling variance by replacing δ_j with the estimate of the mean effect size does not seem to be recommended, whereas the use of the observed effect sizes may be a better choice.

Empirical Bayes effect size estimates (Rubin, 1981) are optimally weighted combinations of these two kinds of estimates, with weights depending on the sampling variation (and thus on the study sizes) and on the variance between study effect sizes:

$$d_{EB} = \lambda_j d_j + (1 - \lambda_j) d_+, \tag{9}$$

where

$$\lambda_j = \hat{\sigma}^2(\delta) \Big/ \left[\hat{\sigma}^2(\delta) + \hat{\sigma}^2(d_j \mid \delta_j)\right].$$

In Equation 9, it can be seen that if the (estimated) true variance is large, less weight will be given to the estimate of the overall effect size. Indeed, the more similar the effect sizes being estimated, the better the estimate of the mean effect size functions as an estimate of the effect size in a certain study. Not only the variance between effect sizes, but also the study sizes are important in choosing an estimate of the individual effect sizes: For a large study, the effect size estimate of that study is relatively reliable, and there is less reason to lean on the estimates from other studies. Therefore, in large studies, a relatively large weight is given to the observed effect size from that study. Although empirical Bayes estimates are not unbiased, the *MSE* is reduced.

Therefore, using the empirical Bayes estimates for estimating the sampling variances seems to be an interesting

alternative to using the weighted mean of effect sizes. Because the empirical Bayes estimates are a result of the meta-analysis, an iterative procedure can be followed, in which the sampling variance estimates and the mean effect size estimate are updated in turn. In cases in which the (estimated) variance between effect sizes is zero, the empirical Bayes estimates are equal to the mean effect size estimate, making this procedure equivalent to the iterative procedure outlined by Hedges and Olkin (1985) for homogeneous studies. In the other extreme situation, in cases in which the estimated variance and/or the study sizes tend to infinity, the empirical Bayes estimates are equal to the observed effect size estimates, making this procedure equivalent to the ordinary procedure, replacing the true effect sizes by the observed ones.

In the following, the results of a small simulation study will be reported. The first purpose of the study is to explore further the extent of the problem of estimating the overall effect size, using the ordinary procedure. Because the unbiasedness is often of less importance than the *MSE*, we will also take the *MSE* into account. A second purpose is to compare the ordinary procedure, the iterative procedure of Hedges and Olkin (1985), and the alternative procedure in which empirical Bayes estimates are used.

A Simulation Study

Meta-analytic datasets were simulated with the following characteristics: group sizes ($n_E = n_C = n$) approximately equal to 10, 25, or 50; number of studies (k) equal to 5, 10, or 50; (mean) true effect sizes, δ, equal to 0.5, 1, or 1.5; and finally, variance between effect sizes, $\sigma^2(\delta)$, equal to 0, 0.1, or 0.2. These values are representative for results that have often occurred in practice or have been used in other simulation studies (e.g., Hedges, Cooper, & Bushman, 1992; Hedges & Vevea, 1998). For each of these 81 combinations, 5,000 meta-analytic datasets were simulated using MLwiN (Goldstein et al., 1998). For each study in each dataset, an effect size δ_j was drawn out of a normal distribution with a mean of δ and a variance of $\sigma^2(\delta)$. To define the group sizes in this study, a value was drawn out of a uniform distribution with bounds of $0.8n$ and $1.2n$, before being rounded. For both groups, this number of observations was drawn from the control condition population distribution, $N(-\delta_j/2, 1)$, and from the experimental condition population distribution, $N(\delta_j/2, 1)$. The "observed" effect size for the study was calculated on the basis of these raw data in the usual way (using Equations 3 and 4).

For each dataset, we used an FEM to estimate the mean of the true effects of the studies included in the dataset, as well as an REM to estimate the mean of the population of true effects from which the true effects were sampled. For both the fixed effects and the random effects procedure, these parameters were estimated using the RIGLS algorithm implemented in MLwiN, resulting in restricted maximum likelihood estimates for the unknown parameters. The estimate of δ was a weighted mean of the observed effect sizes. The weights were estimated in three ways: with the observed effect sizes (d), the weighted average of the observed effect sizes (d_+), and the empirical Bayes estimates of the true effect sizes (d_{EB}). For the approaches in which the overall effect size estimate or the empirical Bayes estimates were used to estimate the weights, we used 10 iterations. To evaluate the results, we estimated the bias of the estimates for each combination, using the estimates for the 5,000 meta-analytic datasets in this combination. This was done by averaging the estimation errors, which are the differences between the estimates and the true parameter values. Note that in cases in which an FEM was used, the parameter that was estimated was the mean of the true effects of the studies included in the dataset. In cases in which an REM was used, the estimated parameter was the mean of the population of true effects. Moreover, we estimated the variance of the estimation errors and the *MSE* around the true parameter value. The results are found in Table 1 (FEM) and Table 2 (REM). To obtain parsimonious tables, we report only the estimated bias and the *MSE* and leave out the results for the intermediate effect size ($\delta = 1$) and the variance [$\sigma^2(\delta) = 0.1$]. Although not reported, the variance estimates can easily be obtained by subtracting the squared bias estimate from the estimated *MSE*, using the following equation:

$$\text{variance} = MSE - (\text{bias})^2. \quad (10)$$

The ordinary procedure. Let us look first at the bias in estimating the mean δ for the FEM if the observed effect sizes are used to estimate the sampling variances. As was expected, the bias is smaller for increasing n. Moreover, the bias increases with increasing true variance and for larger true mean δs. Unexpectedly, the bias seems to increase slightly for an increasing number of studies. This can be explained by noting that the larger the number of studies in a meta-analytic dataset, the more unlikely it is that, in all the studies, similar effect sizes are reported, whereas it is especially in cases in which dissimilar results are combined that the bias is visible. This is most obvious in the extreme case in which there is only one study in a meta- analytic dataset: The effect size estimate of a single study is unbiased.

For the REM, the bias is smaller than that for the FEM, as we expected, because taking into account the variation between true effect sizes equalizes the weights. This is especially true if this true variance is large. Moreover, k, n, and δ seem to have an effect on the bias similar to that for the FEM. Although the effect of the variance is in the same direction as that in the FEM, the effect is strongly reduced: As in the FEM, an increased variance means an increased difference in the estimated sampling variances and, thus, an increase of the bias, but the effect on the weights is largely compensated for by the equalizing effect of the true variance on the weights.

As was expected, the variance of the estimation errors decreases with increasing n and increases slightly with increasing mean δ (this is most easily seen in Table 3, although after some calculation, the values in Table 3 could be deduced from Tables 1 and 2). As a result, the *MSE* is af-

Table 1
Bias (×10,000) and *MSE* (×10,000) for the Fixed Effects Model When the Mean Effect Size Is Estimated, With Weights Based on the Observed Effect Sizes (d), the Weighted Average of the Observed Effect Sizes (d_+), or the Empirical Bayes Estimates (d_{EB})

			Variance 0			Variance 0.2		
δ	n	k	d	d_+	d_{EB}	d	d_+	d_{EB}
				Bias				
0.5	10	5	−214	−14	−65	−389	−28	−211
		10	−222	2	−33	−379	28	−155
		50	−248	−6	−16	−439	−8	−188
	25	5	−70	11	−10	−267	−18	−193
		10	−92	−1	−17	−279	1	−194
		50	−99	−2	−8	−309	−10	−217
	50	5	−25	−16	−26	−205	9	−168
		10	−30	15	7	−245	11	−203
		50	−44	0	−4	−263	−6	−219
1.5	10	5	−568	47	−111	−1,027	−15	−498
		10	−712	38	−138	−1,133	−14	−474
		50	−734	6	−28	−1,222	−9	−463
	25	5	−244	−1	−67	−664	−12	−445
		10	−285	−17	−63	−739	6	−478
		50	−295	0	−19	−810	−2	−527
	50	5	−120	1	−31	−545	0	−432
		10	−133	2	−22	−625	6	−496
		50	−148	−1	−12	−671	0	−533
				MSE				
0.5	10	5	402	435	424	407	453	424
		10	204	222	218	202	222	206
		50	44	42	42	58	46	46
	25	5	163	169	167	167	173	167
		10	78	81	80	88	88	86
		50	17	16	16	25	17	21
	50	5	85	86	85	89	89	88
		10	42	43	42	49	46	48
		50	8	8	8	15	9	13
1.5	10	5	514	555	525	584	568	536
		10	295	283	273	365	279	281
		50	102	56	54	197	57	75
	25	5	208	213	208	259	224	240
		10	110	107	105	157	107	128
		50	29	21	21	87	22	50
	50	5	104	104	104	143	110	133
		10	53	53	53	95	54	81
		50	12	10	10	57	11	40

fected in a similar way (Tables 1 and 2). If an REM is used, increasing the true between-studies variance inflates the variance of the estimates of the overall effect and, consequently, also the *MSE* (Tables 2 and 3). If an FEM is used, the variance of the estimation errors is hardly affected (Table 3). Indeed, when using an FEM, one estimates the mean of the true effect sizes of the specific studies included in the dataset, and the deviation of the estimate from this mean is affected only by the sampling variance within studies. However, as we saw before, increasing the true between-studies variance means an increased bias, especially for a large δ, resulting in an increased *MSE* (Table 1).

With increasing k, the *MSE* is affected more by the decrease in the variance than by the increase in bias, resulting in a decreasing *MSE*.

Although in general, the bias is smaller if an REM is used, the variance and the *MSE* are larger, unless the true between-studies variance is zero. In the REM, the variation of the observed effect sizes and of the weighted means around the true parameter values is indeed affected not only by sampling variation, but also by the variation in true effect sizes. A reduced *MSE* cannot, however, be used as an argument for using an FEM. Rather, the choice of the model must depend on the research interests. When using an FEM, the researcher is interested in the mean of the effects investigated in these specific studies; when an REM is used, the main focus is on the mean effect of the population from which the studies are regarded as a random sample.

Finally, note that the contribution of the bias to the *MSE* is often small. For instance, for the FEM, the bias is −0.0214

Table 2
Bias (× 10,000) and *MSE* (× 10,000) for the Random Effects Model When the Mean Effect Size Is Estimated, With Weights Based on the Observed Effect Sizes (d), the Weighted Average of the Observed Effect Sizes (d_+), or the Empirical Bayes Estimates (d_{EB})

			Variance 0			Variance 0.2		
δ	n	k	d	d_+	d_{EB}	d	d_+	d_{EB}
				Bias				
0.5	10	5	−165	− 14	−43	−175	14	−58
		10	−190	3	−27	−208	30	−60
		50	− 239	−6	−26	−275	−7	−120
	25	5	− 49	11	−1	−129	−51	−95
		10	− 76	−1	−12	−85	7	−49
		50	− 93	−2	−9	−119	−17	−83
	50	5	−45	−16	−21	−28	12	−16
		10	−23	15	10	−64	−18	−53
		50	−45	0	−3	−60	−9	−49
1.5	10	5	−414	50	−44	−548	10	−192
		10	−619	35	−125	−688	3	−242
		50	−702	7	65	−813	−13	−324
	25	5	−181	−1	−36	−304	−78	−197
		10	−241	−16	−48	−253	24	−130
		50	−276	0	−22	−317	−12	−199
	50	5	−88	2	−15	−84	35	−44
		10	−111	2	−14	−140	−1	−100
		50	−138	−1	−11	−142	9	−106
				MSE				
0.5	10	5	410	436	429	787	852	825
		10	207	223	219	378	414	398
		50	44	42	42	83	85	82
	25	5	165	170	168	546	562	553
		10	79	81	80	273	283	276
		50	17	16	16	57	58	57
	50	5	85	86	86	467	475	469
		10	42	43	43	228	242	239
		50	8	8	8	47	47	47
1.5	10	5	526	557	538	894	940	908
		10	291	285	274	492	492	475
		50	98	56	54	153	98	103
	25	5	209	213	210	604	614	608
		10	109	107	106	298	303	299
		50	28	21	21	69	62	64
	50	5	104	105	104	502	509	504
		10	53	53	53	253	255	253
		50	12	10	10	52	51	51

if the meta-analytic dataset consists of five studies with group sizes equal to 10 and true overall effect size equal to 0.5. This means that from the *MSE* of 0.0482, only 0.0005 ($=-0.0214^2$) is due to bias (Equation 12). The bias will be relatively large, as compared with the *MSE*, only when all three of the following are large: k, the true variance, and δ.

Comparing the three approaches. In Tables 1 and 2, it can be seen that the estimates of the bias for the approach in which d_+ is used are small and not systematic. This is true for FEMs as well as for REMs. This corresponds with our expectations that using d_+ results in an elimination of the bias. The results of the empirical Bayes approach are situated in between. Using the empirical Bayes estimates thus reduces the bias, without, however, eliminating the bias.

Concerning the variances of the estimates (which are not reported here), using d_+ generally results in a larger variance of the estimates, as compared with the ordinary procedure. This seems to be true for all combinations and for REM as well as for FEMs. The results of the empirical Bayes procedure are again situated between the two other procedures. For large k, however, the variances in the estimates of the three approaches are comparable. The positive effect on the variance of the estimates of increasing k, indeed, is more pronounced in the d_+ procedure and weakest in the ordinary procedure. The effect of n, the true variance, and δ on the variance of the estimates is similar

Table 3
Variance (× 10,000) of the Estimation Errors for the Fixed Effects Model (FEM) and the Random Effects Model (REM) Using the Observed Effect Sizes (d)

	Variance 0		Variance 0.1		Variance 0.2		Overall	
δ	FEM	REM	FEM	REM	FEM	REM	FEM	REM
0.5	116	117	116	219	122	319	118	219
1	130	132	142	237	162	349	145	237
1.5	159	159	182	269	216	368	186	265
Overall	135	136	147	242	167	344	149	240

in the three approaches. These results can again be explained by the dependence of the weights and the effect sizes. If each study's estimate of d is used to compute the weights, outlying positive or negative observed effects sizes receive a relatively low weight and, hence, do not have a large influence on the results. If, on the contrary, the mean effect size is used to calculate the weights, outlying effect sizes are weighted as much as moderate effect sizes. Thus, if the mean effect size is used, the effect of outliers can be considerable, especially if the number of effect sizes is small, explaining the larger variance in the estimates.

Probably the most important criterion in the comparison is the *MSE*. Although using d_+ generally results in smaller bias and using d in smaller variance of the estimates, the picture is somewhat more complicated for the *MSE*. If the meta-analytic dataset consists of a small number of studies, the *MSE* is often smallest if the observed effect sizes are used to estimate the sampling variances and largest if the d_+ is used. We saw that in the ordinary procedure (and to a smaller degree, in the empirical Bayes procedure), the bias increases with increasing k. Meanwhile, in the ordinary procedure (and to a smaller degree, in the empirical

Table 4
Mean Square Error (×10,000) for the Fixed Effects Model (FEM) and the Random Effects Model (REM) When the Mean Effect Size Is Estimated, With Weights Based on the Observed Effect Sizes (d), the Weighted Average of the Observed Effect Sizes (d_+), or the Empirical Bayes Estimates (d_{EB})

		FEM			REM		
	k	d	d_+	d_{EB}	d	d_+	d_{EB}
$n = 10$	5	469	497	470	653	694	674
	10	260	249	241	336	349	337
	50	95	50	52	93	70	70
$n = 25$	5	195	192	192	380	389	384
	10	106	96	99	192	196	193
	50	37	19	25	42	39	40
$n = 50$	5	103	96	100	291	295	292
	10	58	49	54	147	149	148
	50	21	10	16	30	29	29
$\delta = 0.5$	5	217	233	224	411	430	422
	10	110	116	112	203	214	209
	50	28	23	24	43	43	42
$\delta = 1$	5	249	256	249	434	453	444
	10	138	129	129	225	232	226
	50	47	26	31	53	46	46
$\delta = 1.5$	5	302	296	289	479	495	484
	10	177	149	152	248	249	243
	50	79	29	39	69	49	50
$\sigma^2(\delta) = 0$	5	244	258	250	247	259	254
	10	128	129	126	127	130	127
	50	34	25	25	33	25	25
$\sigma^2(\delta) = 0.1$	5	252	260	250	444	462	453
	10	139	132	129	225	231	226
	50	49	26	29	56	48	46
$\sigma^2(\delta) = 0.2$	5	271	267	263	632	658	644
	10	156	133	138	323	332	325
	50	70	27	40	76	67	67
Overall	5	256	262	254	441	459	450
	10	142	131	131	225	231	226
	50	51	26	31	55	46	46
		149	140	139	240	246	241

Bayes procedure), the variance decreases more slowly than in the d_+ procedure. The result is that whereas for a small number of studies the *MSE* is smallest if the observed effect sizes are used, for a large number of studies the d_+ procedure gives the best results. In between, for a moderate number of studies, one can often profit from the empirical Bayes procedure. This trend can be seen in Table 4.

A complicating factor is that the number of studies in the meta-analytic dataset for which each of the three procedures performs the best depends on other characteristics of the meta-analytic dataset. In Table 4, it can be seen that the number of studies for which the empirical Bayes or the d_+ approach is better than the ordinary approach is smaller if n is larger: If, for instance, an FEM is used when $n = 10$, the ordinary procedure performs the best if $k = 5$, closely followed by the empirical Bayes approach; for $k = 10$, the empirical Bayes approach is the best; for $k = 50$, the d_+ procedure has become the best procedure. When $n = 50$, on the contrary, the d_+ procedure yields the best results, even for a very small number of studies. The effect of n is similar for an REM. Similarly, it can be concluded, again for both FEMs and REMs, that the number of studies for which the empirical Bayes or the d_+ approach is better than the ordinary approach is smaller if δ is larger. In addition, we found that the number of studies for which the empirical Bayes or the d_+ approach is better than the ordinary approach is smaller if the variance is larger. This, however, seems to be true only for the FEM. Finally, we conclude that the number of studies for which the empirical Bayes or the d_+ approach is better than the ordinary approach is smaller if an FEM is used.

Conclusions

In the article, we focused on the problem of estimating the mean effect size. As has been pointed out before (e.g., Hedges, 1983), the precision-weighted average of observed effect sizes resulting from homogeneous studies, with optimal weights estimated using the observed effect sizes, gives a biased estimate of the mean effect. Although we found that the problem of bias is even larger when the studies are heterogeneous and that the bias does not disappear with increasing k, but even increases, the results of our simulation study suggest that the bias is often relatively small, taking into consideration the variance of the estimates. Only when k, the true variance, and δ are large is a substantial part of the *MSE* due to bias.

In the article, we have proposed an alternative approach: The overall effect size is still estimated using a precision-weighted mean of the observed effect sizes, but this time, empirical Bayes estimates of the true effect sizes are used to estimate the precision. The approach is compared with the ordinary approach and with an iterative approach previously presented by Hedges and Olkin (1985). Although the bias disappears when the iterative approach proposed by Hedges and Olkin is used and the variance of the estimates is usually smallest when the ordinary approach is used, there is no uniformly best approach regarding the *MSE*.

An interesting result is that the iterative approach of Hedges and Olkin (1985), using the overall effect size estimate to estimate the weights, may even exacerbate the problem by increasing the *MSE*. Still, this procedure often means an improvement if the number of studies is large. This is true even when study results are heterogeneous and the overall effect size estimate, therefore, is only a rough estimate of the individual effect sizes. In general, if the meta-analytic dataset consists of a small number of studies, the ordinary approach usually performs better; for a moderately sized dataset, the empirical Bayes approach often yields the best results; for a large meta-analytic dataset, the iterative approach of Hedges and Olkin usually can be recommended. This is true for both REMs and FEMs. However, the sizes of the meta-analytic dataset for which each of the three procedures performs the best depend on other characteristics of the meta-analytic dataset. The number of studies for which the empirical Bayes or the d_+ approach is better than the ordinary approach is larger if δ is small, if n is small, and if an REM is used. For the FEM, this number gets smaller with increasing true variance.

REFERENCES

COHEN, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

DERSIMONIAN, R., & LAIRD, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177-188.

GLASS, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, **5**, 3-8.

GLASS, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, **5**, 351-379.

GLASS, G. V., MACGAW, B., & SMITH, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

GOLDSTEIN, H., RASBASH, J., PLEWIS, I., DRAPER, D., BROWNE, W., YANG, M., WOODHOUSE, G., & HEALY, M. (1998). *A user's guide to MLwiN*. London: University of London, Multilevel Models Project.

HEDGES, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**, 107-128.

HEDGES, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, **92**, 490-499.

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

HEDGES, L. V., COOPER, H., & BUSHMAN, B. J. (1992). Testing the null hypothesis in meta-analysis: A comparison of combined probability and confidence interval procedures. *Psychological Bulletin*, **111**, 188-194.

HEDGES, L. V., & OLKIN, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

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

HUNTER, J. E., & SCHMIDT, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. London: Sage.

HUNTER, J. E., SCHMIDT, F. L., & JACKSON, G. B. (1982). *Meta-analysis: Cumulating findings across research*. Beverly Hills, CA: Sage.

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*. Newbury Park, CA: Sage.

RUBIN, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, **6**, 377-400.

VAN DEN NOORTGATE, W., & ONGHENA, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational & Psychological Measurement*, **63**, 765-790.

(Manuscript received March 5, 2002;
revision accepted for publication January 14, 2003.)