Behavior Research Methods
2005, 37 (1), 11-22

Parametric and nonparametric bootstrap methods for meta-analysis

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

In a meta-analysis, the unknown parameters are often estimated using maximum likelihood, and inferences are based on asymptotic theory. It is assumed that, conditional on study characteristics included in the model, the between-study distribution and the sampling distributions of the effect sizes are normal. In practice, however, samples are finite, and the normality assumption may be violated, possibly resulting in biased estimates and inappropriate standard errors. In this article, we propose two parametric and two nonparametric bootstrap methods that can be used to adjust the results of maximum likelihood estimation in meta-analysis and illustrate them with empirical data. A simulation study, with raw data drawn from normal distributions, reveals that the parametric bootstrap methods and one of the nonparametric methods are generally superior to the ordinary maximum likelihood approach but suffer from a bias/precision tradeoff. We recommend using one of these bootstrap methods, but without applying the bias correction.

In a meta-analysis, the results of a set of studies are integrated quantitatively. Raudenbush and Bryk (1985) showed that a meta-analysis can be considered as a multilevel analysis with participants *nested* within studies and that multilevel models or hierarchical linear models are useful in meta-analysis. An important advantage of using hierarchical linear models for meta-analysis is their flexibility, allowing, for instance, the performance of multivariate meta-analyses, the combining of effect sizes and raw data in one analysis, or the modeling of dependencies between studies by including higher levels (Goldstein, Yang, Omar, Turner, & Thompson, 2000).

Parameters of hierarchical linear models are often estimated using maximum likelihood, and inferences are based on asymptotic theory. Unfortunately, for a meta-analysis, this may yield biased parameter estimates and standard errors, as we shall discuss further. A flexible technique with which to obtain bias-corrected parameter estimates and corresponding standard errors is the bootstrap (Efron, 1982). The use of the bootstrap for hierarchical linear models has been discussed before (e.g., Carpenter, Goldstein, & Rasbash, 1999; Laird & Louis, 1987, 1989; Meijer, van der Leeden, & Busing, 1995), but its implementation for meta-analysis has received little attention. Nevertheless, the application of the bootstrap for meta-analysis is not straightforward, because, as we will discuss below, usually not all raw data are available. Moreover, because of the computing intensity of bootstrap methods, until recently, simulation studies evaluating bootstrap methods were practically infeasible, and little is known about their performance in multilevel analysis (Davison & Hinkley, 1997; Rasbash et al., 2000), let alone in a multilevel meta-analysis.

In the following, we will first describe the use of hierarchical linear models for meta-analysis and problems associated with maximum likelihood estimation. Then we will apply the idea of the bootstrap to meta-analysis. Several bootstrap methods will be proposed and illustrated with empirical data. Finally, the bootstrap methods for meta-analysis will be evaluated with a simulation study.

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

Note—This article was accepted by the previous editor, Jonathan Vaughan.

Using Hierarchical Linear Models for Meta-Analysis

In a hierarchical linear model, the scores of the dependent variable are regressed on characteristics of the first-level units. The coefficients of this regression equation may vary over Level 2 units, in which the Level 1 units are *nested*. The variation over Level 2 units is described using new regression equations including Level 2 characteristics as predictors. The coefficients of these second-level equations, in turn, may be described on a third level, and so on.

A hierarchical linear model with two levels and one predictor at each level is defined by the following equations:

Level 1

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \quad (1)$$

and

Level 2

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j},$$

with $i = 1, 2, \ldots, n_j$ indicating the Level 1 units, $j = 1, 2, \ldots, J$ indicating the Level 2 units, X and Z being characteristics of the Level 1 and Level 2 units, respectively, and the es and us being residuals on Level 1 and Level 2, respectively.

This kind of model can, for example, be applied in educational research to model data from pupils (Level 1 units) grouped in schools (Level 2 units). Pupils' scores on a math test may be regressed on the pupils' age. The regression weight of this Level 1 predictor and the intercept possibly vary over schools, indicating that the effect of age and the performance of pupils after correction for age depend on the school. A school characteristic (e.g., school size) may be included to explain the between-school variation in the effects of age and in the base level.

The hierarchical linear model (Equation 1) usually cannot be used immediately in meta-analyses. A meta-analysis differs from an ordinary multilevel analysis in that the data on the first level (i.e., the scores of the study participants) are usually summarized for each Level 2 unit (i.e., per study)—for instance, by using summary statistics, such as group means and standard deviations. Moreover, the measurement scale of the dependent variable may depend on the study. Therefore, in a meta-analysis, data are usually converted to a standardized effect size measure (e.g., a correlation coefficient or a standardized mean difference) for each study separately, and these effect size measures are combined. The basic meta-analytic hierarchical linear model is a two-level model, with participants (Level 1) nested within studies (Level 2; Raudenbush & Bryk, 2002):

$$d_j = \quad_j + e_j \qquad \text{(Level 1)}$$

and

$$\quad_j = \quad_0 + \sum_{s=1}^{S} \quad_s Z_{sj} + u_j, \qquad \text{(Level 2)}$$

or combined,

$$d_j = \quad_0 + \sum_{s=1}^{S} \quad_s Z_{sj} + u_j + e_j, \qquad (2)$$

with $j = 1, 2, 3, \ldots, k$ indicating the study, d_j the observed effect size in study j, $\quad_j$ the "true" effect size of study j, e_j the sampling error associated with d_j as an estimate of $\quad_j$, $Z_1, \ldots, Z_S$ study characteristics predicting the effect sizes, $\quad_0, \ldots, \quad_S$ regression coefficients, and u_j a random Level 2 residual.

Whereas in an ordinary multilevel analysis, we have a residual for each Level 1 unit, in a meta-analysis we have on Level 1 only one residual for each Level 2 unit. This means that without additional information, the Level 2 variance (or between-study variance) cannot be distinguished from the Level 1 variance (or within-study or sampling variance). Nevertheless, the sampling variances of the observed effect sizes can often be estimated using the reported study results—for example, results based on the study sizes. Raudenbush and Bryk (2002), therefore, call a meta-analysis a "variance known problem" and propose constraining, in the meta-analysis, the Level 1 variances to their estimated values. For that purpose, Equation 2 is rewritten by including the (estimated) standard errors ($_j$) of the observed effect sizes as a predictor with a random (or study-specific) coefficient:

$$d_j = \quad_0 + \sum_{s=1}^{S} \quad_s Z_{sj} + u_j + \left(r_j \lozenge \hat{} \quad_j\right). \qquad (3)$$

The Level 1 variance—the sampling variance of d_j given $_j$—thus equals the variance of $r_j \cdot \hat{} \quad_j$. Since $\hat{} \quad_j$ is constant for study j, this variance is equal to $\hat{}^2_j$ times the variance of r_j. By constraining the variance of the random coefficient r_j to 1, the Level 1 variance therefore is constrained to $\hat{}^2_j$, the estimated sampling variance of d_j.

The researcher is typically interested in the second-level coefficients (the s, also-called *fixed* coefficients) and in the variance component on the second level, although sometimes the individual study effects (the $_j$s) may also be of interest. The parameters of hierarchical linear models are usually estimated using maximum likelihood, assuming a univariate or multivariate normal distribution for the residuals on each level and independence of the residuals from different levels. For the meta-analytic model of Equation 2, this means that the Level 1 and the Level 2 residuals are independently distributed, with e_j $N(0, \quad^2_j)$ and u_j $N(0, \quad^2)$.

Unless data are perfectly balanced, the maximum likelihood estimates cannot be derived in closed form but are estimated using iterative procedures, updating consecutively the estimates of the fixed coefficients and of the variance components. The exact sampling distributions of the maximum likelihood estimates are unknown. Inferences or interval estimates are, therefore, often based on the asymptotic normality of the sampling distributions of maximum likelihood estimates.

Several problems are associated with the use of maximum likelihood for hierarchical linear models in general or for meta-analysis in particular. First, the estimated standard errors and the assumption of normal sampling distributions for the maximum likelihood estimates apply only asymptotically. For small samples, the sampling distributions for the fixed coefficients are closely approximated by t distributions, but for the variance components the sampling distributions are skewed to an unknown degree, and the estimates are biased (Meijer et al., 1995; Raudenbush & Bryk, 2002; Van den Noortgate & Onghena, 2003b).

Second, the estimates of the fixed coefficients and the corresponding standard errors depend on the point estimates of the variance components. This means that inferences for the fixed coefficients are conditional on the accuracy of these point estimates. Unfortunately, estimated standard errors and, therefore, interval estimates and significance tests for the fixed coefficients fail to take into account this additional uncertainty about the estimates of the variance components (Laird & Louis, 1987).

Third, the use of the maximum likelihood theory can yield misleading results if the assumption of (multivariate) normality on each level is violated (Raudenbush & Bryk,

2002). Van den Noortgate and Onghena (2003b) show that the results of a meta-analysis are relatively robust for a violation of the normality of the Level 2 residuals, but more research is needed about the influence of a violation of the assumption that the sampling distribution of the effect sizes (Level 1) is normal.

Finally, although the maximum likelihood method, in principle, yields unbiased estimates of the fixed coefficients, the estimate of the mean population effect size may be biased (Van den Noortgate & Onghena, 2003a). The mean population effect size is estimated by averaging the observed effect sizes weighted by the inverse of their (estimated) variance around the mean effect size. Sometimes, this variation and, thus, the weight of an observed effect size depend on the observed effect size itself. For example, to express the difference between an experimental group and a control group, the difference between the sample means divided by the square root of the pooled variance is often used, assuming a common population variance. The sampling variance of this standardized mean difference equals (Hedges, 1981):

$$^{2}(d_j) \quad \frac{N}{n_E n_C} + \frac{^{2}_{j}}{2N}, \qquad (4)$$

with n_E and n_C equal to the size of the experimental and the control groups, respectively, and $N = n_E + n_C$. For an observed standardized mean difference, the sampling variance is estimated by replacing $_j$ from Equation 4 with the observed value d_j. Equation 4 indicates that the larger the (observed) standardized mean difference, the larger its estimated standard error. This means that in a meta-analysis, larger observed effect sizes get smaller weights. Although the sample standardized mean difference is an (almost) unbiased estimate of the corresponding population parameter and the unweighted mean of the observed effect sizes, therefore, is an unbiased estimate of the mean population effect size, the weighted mean of the observed effect sizes, with smaller weights for larger effect sizes, is a negatively biased estimate of a positive mean effect size.

In the following, we will discuss the bootstrap method for meta-analysis and evaluate whether the bootstrap yields improved parameter estimates and corresponding standard errors.

The Bootstrap for Meta-Analysis

The bootstrap for hierarchical linear models is based on the idea that if the fixed parameters and the distributions of the residuals were known, the sampling distribution of a parameter estimate could be approximated by simulating a large number of samples (B)—say, 5,000—and estimating the parameter of interest for each sample. In general, the larger the number of samples drawn, the better the distribution of the B estimates resembles the exact sampling distribution. Of course, the fixed parameters and the distributions of the residuals are unknown. In the bootstrap, these population characteristics are estimated using the empirical data, and the sampling distribution of the parameter of interest is approximated by simulating data from the estimated population distributions (Efron, 1982).

The bootstrap is often used to correct parameter estimates for bias. Since the bootstrap estimates are related to the initial estimates in the same way as the initial estimates are related to the population parameters, one can estimate the bias and correct the initial estimates (Efron, 1982). If, for instance, it is found that for the bootstrap samples, the estimates of a specific parameter are, in general, higher than the initial estimate, one can assume that the initial estimate also overestimates the population value to the same degree and can correct this estimate by subtracting the estimated bias.

Second, the distribution of the B estimates can be used for significance testing or confidence intervals. For instance, assuming that the sampling distribution of a specific parameter estimate is normal, parametric confidence intervals can be obtained using the standard deviation of the B estimates as an estimate of the standard error of estimation. Alternatively, one can use the *percentile method* and construct nonparametric confidence intervals by calculating percentiles from the bootstrap estimates (Efron, 1982). More complex methods for constructing confidence intervals have been proposed in general bootstrap theory (see, e.g., Efron & Tibshirani, 1993), but a discussion of these methods is beyond the scope of this article.

Three remarks must be made regarding the correction for bias. First, although negative variance component estimates must be constrained to zero in order to draw bootstrap samples, an appropriate bootstrap standard error is obtained by allowing negative variance estimates for the bootstrap samples. The bias is estimated by comparing the mean of the unconstrained variance estimates from the bootstrap samples with the initial constrained variance estimate and is used to correct the initial unconstrained variance estimate. Negative quantiles or negative means of the variance estimates are constrained to zero afterward. Second, in cases in which there is evidence that the bias depends on the parameter value itself, an iterative bias correction procedure may be useful (Goldstein, 1996; Kuk, 1995). Finally, Rasbash et al. (2000) suggested *scaling* the bootstrap estimates of the standard errors by multiplying them by the ratio of the bias-corrected and the initial parameter estimates, since the corrected and the uncorrected parameter estimates may show different variances.

In the following, we will discuss the simulation of bootstrap samples in meta-analysis. Since, in meta-analyses, there is usually only one Level 1 residual for each study—because the data for each study are typically summarized with an effect size measure—and the sampling variance of this residual is (assumed) known, bootstrap samples cannot be drawn in the same way as that for typical hierarchical linear models. We will describe four possible methods and give them names that are consistent with names given in the literature on bootstrapping in multilevel analysis (e.g., Busing, Meijer, & van der Leeden, 1994; Meijer et al., 1995). The first two methods, the *effect size*

bootstrap and the *raw data bootstrap,* are parametric bootstrap methods, since assumptions are made about the distributions of the data. The *error bootstrap* and the *cases bootstrap* are nonparametric bootstrap methods.

The effect size bootstrap. In the parametric effect size bootstrap, bootstrap samples are obtained in four steps.

1. Draw a set of k Level 2 residuals $(u_1^*, u_2^*, \ldots, u_k^*)$ from the parametrically estimated distribution of Level 2 residuals—for example, from $N(0, \hat{\ }^2)$.

2. Derive a set of k *true* study effect sizes ($^*_1$, $^*_2$, . . . , $^*_k$) by adding these Level 2 residuals to the fixed part of the meta-analytic model, using the initial (maximum likelihood) estimates for the fixed parameters:

$$^*_j = \hat{\ }_0 + \sum_{s=1}^{S} \hat{\ }_s W_{sj} + u_j^*. \quad (5)$$

3. Draw a set of k Level 1 residuals $(e_1^*, e_2^*, \ldots, e_k^*)$ from the parametrically estimated sampling distributions—for example, from $N(0, \hat{\ }^2_j)$, with $\hat{\ }^2_j$ estimated using Equation 4, based on the $^*_j$ derived in the second step.

4. Derive the bootstrap sample of *observed* effect sizes, $(d_1^*, d_2^*, \ldots, d_k^*)$, by adding these Level 1 residuals to the *true* study effect sizes obtained in the second step:

$$d_j^* = {}^*_j + e_j^*. \quad (6)$$

Some remarks must be made. First, the effect size bootstrap assumes that the model is correctly specified. One also assumes that the population distributions of the residuals are of a known family—for instance, normal.

Second, the sampling variance (^2_j) of the standardized mean difference depends on the true effect size $_j$ (Equation 4). For some effect size measures (e.g., the log odds ratio), effect size and sampling variance are independent. In this case, a bootstrap sample can be drawn in one step, assuming normal distributions for the residuals on both levels (see, e.g., Laird & Louis, 1989; Smith, Caudill, Steinberg, & Thacker, 1995):

$$d_j^* \quad N\left(\hat{\ }_0 + \sum_{s=1}^{S} \hat{\ }_s Z_{sj}, \hat{\ }^2_j + \hat{\ }^2\right). \quad (7)$$

Finally, if the study characteristics included in the Level 2 model (the Zs from Equation 2) are regarded as fixed variables, each bootstrap sample should have exactly the same values for these variables as those in the initial set of studies (Busing et al., 1994). If a study characteristic Z is regarded as random, bootstrap samples $(Z_1^*, Z_2^*, \ldots, Z_k^*)$ are randomly drawn, before using Equation 5 to obtain samples of true effect sizes. If the distribution of Z is not specified, nonparametric samples are drawn from the set of observed values. Alternatively, parametric samples can be drawn, assuming, for instance, that Z is normally distributed (Busing et al., 1994).

The raw data bootstrap. When the ordinary maximum likelihood is used for meta-analysis, the sampling distribution of the effect size measure is assumed to be normal. Also, for obtaining bootstrap samples in the effect size bootstrap described above, the sampling distribution of the effect size is assumed to be normal (or possibly, to belong to another known family). Yet, for some commonly used effect size measures (e.g., the standardized mean difference, the correlation coefficient, the log odds ratio, etc.), the sampling distribution is only approximately normal. For instance, Hedges (1981) showed that if the scores of two populations are independently normally distributed with a common variance, the exact sampling distribution of the standardized mean difference is linearly related to a noncentral t distribution. Under other conditions, the sampling distribution of the standardized mean difference may be difficult or impossible to derive. For some measures of effect size, the sampling distribution is unknown.

To free the assumption of a normal sampling distribution of the effect size, we propose an alternative parametric bootstrap method for meta-analysis, which differs from the effect size bootstrap in the third and fourth steps for getting bootstrap samples. Bootstrap samples are simulated as follows.

1. Draw a set of k Level 2 residuals $(u_1^*, u_2^*, \ldots, u_k^*)$ from the parametrically estimated distribution—for instance, $N(0, \hat{\ }^2)$.

2. Derive a set of k true study effect sizes ($^*_1$, $^*_2$, . . . , $^*_k$) by adding these residuals to the estimated fixed part.

3. Draw raw data for each study, conditional on the true effect sizes from the preceding step.

4. Use the simulated raw data to calculate the effect size for each study.

Instead of drawing bootstrap samples of observed effect sizes from normal distributions with means and variances equal to the true effect sizes obtained in the second step and the estimated sampling variances, respectively, as is done in Steps 3 and 4 of the effect size bootstrap, raw data are simulated and used to calculate the sample of observed effect sizes $(d_1^*, \ldots, d_k^*)$. Therefore, the preceding method was called the *effect size* bootstrap, and this one the *raw data* bootstrap.

The way the raw data are drawn depends on the effect size measure used. For instance, if the standardized mean difference is used as a measure of effect size to express the difference between two groups, raw data for both groups from study j can be drawn from

$$N\left(\frac{^*_j}{2}, 1\right)$$

and

$$N\left(\frac{^*_j}{2}, 1\right),$$

respectively, assuming normal distributions with a common variance. If the correlation coefficient is used, one could sample raw data for study j from a bivariate normal distribution with zero means, variances equal to 1, and a covariance equal to $_j$. Note that since the raw data are used only to calculate the effect size, they can also be sampled from other population distributions, as long as

the population effect sizes are equal to the effect sizes sampled in the second step of the bootstrap algorithm and the populations correspond to the assumptions one wants to make (e.g., normality). A similar procedure was used by Van den Noortgate and Onghena (2003c), who proposed a parametric bootstrap version of Hedges's (1982) homogeneity test for meta-analysis.

The error bootstrap. In nonparametric bootstrap methods, no assumptions are made about the kind of distribution for the residuals. Instead, the population distribution of the data is approximated by the empirical distribution, and bootstrap samples are obtained by drawing with replacement from the empirical data.

In the nonparametric error bootstrap, the bootstrap samples are simulated in a way similar to that in the effect size bootstrap, but instead of drawing residuals from normal distributions, residuals are drawn with replacement from the set of estimated residuals, as follows.

1. Draw with replacement a set of k Level 2 residuals $(u_1^*, u_2^*, \ldots, u_k^*)$ from the estimated Level 2 residuals $(\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_k)$.

2. Derive a set of k true study effect sizes $(\ _1^*, \ _2^*, \ldots, \ _k^*)$ by adding these residuals to the fixed part of the meta-analytic model, using the initial estimates for the fixed parameters (Equation 5).

3. Draw with replacement a set of k Level 1 residuals $(r_1^*, r_2^*, \ldots, r_k^*)$ from the estimated Level 1 residuals $(\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_k)$ and multiply these residuals by the square roots of the estimated sampling variances of the observed effect sizes, to obtain $(e_1^*, e_2^*, \ldots, e_k^*$; see Equations 2 and 3).

4. Derive the bootstrap sample of effect sizes $(d_1^*, d_2^*, \ldots, d_k^*)$ by adding the Level 1 residuals $(e_1^*, e^*, \ldots, e_k^*)$ to the true study effect sizes obtained in the second step (Equation 6).

The (empirical Bayes) estimates of the residuals for study j are equal to

$$\hat{u}_j = \frac{\hat{\ }^2}{\hat{\ }^2 + \hat{\ }_j^2} \left(d_j - \hat{\ }_0 - \sum_{s=1}^{S} \hat{\ }_s W_{sj} \right)$$

and

$$\hat{r}_j = \frac{\hat{e}_j}{\hat{\ }_j} = \frac{d_j - \hat{u}_j}{\hat{\ }_j}. \qquad (8)$$

Note that if bootstrap samples are constructed by resampling residuals obtained using Equation 8, one could generally expect negatively biased variance estimates, because the estimated residuals are *shrunk* toward zero. Therefore, the residuals must be *reflated*, before being used to obtain bootstrap samples, by multiplying the (estimated) residuals by a constant making the variance of the residuals equal to the initial variance estimates (Carpenter et al., 1999). The estimated r_js are also reflated, so that their variance becomes 1.

Note that multiplying the sample of r_js by the square root of the estimated sampling variances, the $\hat{\ }_j$s, may raise problems if the sampling variance depends on the effect size, as is the case with the standardized mean difference. Analogous to the adaptation described above for the effect size bootstrap, we therefore propose to calculate the expected sampling variances, $\hat{\ }_j^{2*}$s, given the true effect sizes, $\ _j^*$, that are sampled in the second step (using Equation 4), and to multiply the square root of these estimated sampling variances with the resampled and reflated r_js to obtain a sample of Level 1 residuals. Unfortunately, to estimate $\hat{\ }_j^*$, assumptions must typically be made about the population distribution of the raw data, making this procedure not fully nonparametric.

The cases bootstrap. Meijer et al. (1995) have described the cases bootstrap for a two-level hierarchical linear model, drawing units with replacement, together with the corresponding values for the dependent and independent variables. The cases bootstrap can be useful if the predictors are random. In the first step, k *cases* are drawn with replacement from the k Level 2 units (e.g., schools). In the second step, n_j Level 1 units (e.g., pupils) are drawn with replacement for each Level 2 unit drawn in the first step. It is also possible to resample cases on the first or the second level only—for instance, when the Level 2 or Level 1 units cannot be considered random.

Because in a meta-analysis the raw data on the first level are not available, no cases can be drawn of the first level. Instead, in the cases bootstrap for meta-analysis, k studies are drawn with replacement from the set of the k studies, including the observed effect size, the corresponding sampling variance, and the study characteristics.

Finally, we note that combinations are possible. For instance, Zhou, Brizendine, and Pritz (1999) combined the cases and the raw data bootstraps by drawing with replacement k studies and, subsequently, raw data, given the observed effect sizes for these studies.

An Example

To illustrate the bootstrap methods, we reanalyze the data analyzed by Hedges and Olkin (1985, p. 25). The data set includes the results of 10 well-controlled studies evaluating the effects of open-education programs on student creativity by comparing studies in experimental open-classroom schools with those of students from traditional schools. The data set is a part of a much larger data set assembled by Hedges, Giaconia, and Gage (1981), including studies in which the effect of open education on a variety of outcome variables was investigated. The data needed for our analysis are presented in Table 1.

We calculated the standard error associated with each observed effect size, using Equation 4. Next, we analyzed the data with an *empty* meta-analytic model, which is a model without predictors:

$$d_j = \ _0 + u_j + e_j. \qquad (9)$$

Unknown parameters, thus, are the mean effect over studies $(\ _0)$ and the between-study variance $(\ ^2)$. Parameters were estimated with the MLwiN software (Rasbash et al., 2000), using the restricted iterative generalized least squares (RIGLS) algorithm, giving restricted maximum likelihood estimates if residuals are normally dis-

Table 1
Results of Experimental Studies Regarding the Effects of Open Education on Student Creativity (Hedges & Olkin, 1985)

Study	Grade Level	Sample size ($n_E = n_C$)	Standardized Mean Difference (d_j)
1	6	90	−0.583
2	5	40	0.535
3	3	36	0.779
4	3	20	1.052
5	2	22	0.563
6	4	10	0.308
7	8	10	0.081
8	1	10	0.598
9	3	39	−0.178
10	5	50	−0.234

tributed (Goldstein, 1995). The results are given in Table 2 (first column, with initial RIGLS estimates).

The estimate of the mean effect is equal to 0.252, which refers to a rather small effect.[1] Using a (two-sided) Wald test, a comparison of the ratio of the parameter estimate and the corresponding standard error with a standard normal distribution, reveals that this estimate is statistically not significant at the .05 level ($z = 1.41$, $p = .16$). The estimate of the variance between studies equals 0.230, which is, however, statistically not significant when the Wald test is used[2] ($z = 1.63$, $p = .10$).

A visual inspection of Table 1 reveals that the observed difference between the groups is largest for smaller grade levels. To investigate this relation, we included the grade level in the model:

$$d_j = {}_0 + {}_1.(\text{grade})_j + u_j + e_j. \qquad (10)$$

The intercept from Equation 10 indicates the expected effect in Grade 0; the coefficient of the grade indicates how much the expected effect generally increases for each grade. To give a meaningful interpretation to the intercept, we subtracted 1 from the grades. As a result, the intercept can be interpreted as the expected effect in Grade 1. The results are again presented in Table 2 (second column, with initial RIGLS estimates). The expected effect in Grade 1 is large (0.730) and statistically significant ($z = 2.32$, $p = .02$), but the expected effect decreases substantially with the level, although the decrease is statistically not significant at the .05 level ($z = 1.80$, $p = .07$). By taking the grade level into account, the estimated between-study variance decreases approximately 30%.

We also estimated the parameters of the model, using the four bootstrap methods described above. For each method, we drew 10,000 bootstrap samples and again estimated the unknown parameters, using the RIGLS algorithm. The bootstrap estimates of the parameters and the corresponding standard errors are the means and standard deviations of the RIGLS estimates of the 10,000 bootstrap samples. For the effect size bootstrap, samples of true effect sizes ($^*_1$, $^*_2$, . . . , $^*_k$) were obtained by adding residuals drawn from $N(0, 0.230)$ (empty model) or from $N(0, 0.163)$ (model with grade as a predictor) to the estimated fixed part of the model, which is 0.252 for the first model and 0.730 − 0.160∗grade for the second model. Next, Level 1 residuals are drawn from $N(0, \hat{}^2_j)$, with $\hat{}^2_j$ calculated using Equation 4 and based on the true effect sizes drawn in the previous step. In the raw data bootstrap, samples of true effect sizes were first drawn as in the effect size bootstrap. Next, raw data for both groups from study j were drawn from

$$N\left(\frac{^*_j}{2}, 1\right)$$

and

$$N\left(\frac{^*_j}{2}, 1\right),$$

respectively, assuming normal distributions with a common variance, and were used to calculate d_j. In the error bootstrap, residuals were estimated using Equation 8, and samples of residuals were drawn with replacement from the estimated residuals. Adding a sample of Level 2 residuals to the estimated fixed part gave a sample of true effect sizes, which could be used to estimate the sampling variances (Equation 4), with which the Level 1 residuals were multiplied before being added to the true effect sizes to result in a new sample of effect sizes. In the cases bootstrap, samples were obtained by drawing whole studies with replacement from our initial set of 10 studies. All calculations and analyses were done using MLwiN. (MLwiN macros for obtaining the bootstrap estimates for the example can be obtained from the authors.)

If for a specific bootstrap sample the RIGLS algorithm was not converged after 50 iterations, the sample was discarded, and a new sample was drawn. When the parameters of a hierarchical linear model are estimated iteratively, MLwiN sometimes gives the warning that the covariance matrix has gone negative definite. If one chooses to proceed, MLwiN will automatically approximate the covariance matrix by the nearest positive definite matrix, using a singular value decomposition, and convergence may eventually be achieved smoothly. Still, even if convergence is reached, it is recommended that one examine carefully the specification of the model. In the bootstrap procedures implemented in MLwiN for "ordinary" hierarchical linear models, bootstrap samples that lead to a negative definite matrix therefore are ignored (Rasbash et al., 2000). In our analysis, we also ignored bootstrap samples for which this problem occurred.

Table 2 shows that the bootstrap estimates of the parametric bootstrap methods and of the nonparametric error bootstrap are similar and differ from the results of the nonparametric cases bootstrap. For the former methods, the bootstrap estimates of the intercept, the coefficient of grade, and the between-study variance are somewhat closer to zero than the initial RIGLS estimates are. Since the initial RIGLS estimates were used as population values to draw bootstrap samples, a difference between the mean (RIGLS) estimates for the bootstrap samples and the ini-

Table 2
Parameter Estimates (Est, With Standard Errors) for the Open Education Data

	Initial RIGLS		Equation 9								Initial RIGLS		Equation 10					
			ES		RD		Error		Cases				ES		RD		Error		Cases	
Parameter	Est	*SE*	Est	*SE*	Est	*SE*	Est	*SE*	Est	*SE*	Est	*SE*	Est	*SE*	Est	*SE*	Est	*SE*	Est	*SE*
Intercept	0.252	0.179	0.249	0.177	0.247	0.177	0.254	0.177	0.256	0.173	0.730	0.315	0.714	0.312	0.714	0.314	0.707	0.311	0.761	0.284
Effect grade											−0.160	0.089	−0.155	0.089	−0.156	0.088	−0.153	0.089	−0.169	0.087
Between-study variance	0.230	0.141	0.221	0.149	0.221	0.150	0.222	0.130	0.192	0.090	0.163	0.109	0.160	0.124	0.157	0.123	0.161	0.103	0.113	0.062

Note—RIGLS, restricted iterative generalized least squares algorithm; ES, effect size bootstrap; RD, raw data bootstrap; Error, error bootstrap; Cases, cases bootstrap.

tial RIGLS estimates indicates that the RIGLS estimates are biased. For instance, whereas for the empty model, bootstrap samples are sampled from a population with a between-study variance of 0.230, the mean of the sample estimates in the effect size bootstrap is 0.221. The estimated bias thus is −0.009. If the RIGLS algorithm gives negatively biased variance estimates, the initial variance estimate (0.230) will probably also be biased to a comparable degree. Therefore, a bias-corrected estimate of the between-study variance equals 0.230 − (−0.009) = 0.239.

The cases bootstrap suggests that the estimates of the fixed coefficients are too large and that there is a relatively large negative bias in estimating the between-study variance. Bias-corrected estimates, therefore, will be closer to zero than the initial RIGLS estimates for the fixed parameters but more extreme for the variance parameter.

The bootstrap standard errors for the fixed coefficients are equal or slightly smaller than the initial standard error estimates. Only for the cases bootstrap are they substantially smaller. The standard error of the variance component gets somewhat larger when the parametric bootstrap is used, somewhat smaller when the error bootstrap is used, and much smaller when the cases bootstrap is used. The standard errors could be scaled by multiplying them by the ratio of the bias—corrected to the initial estimate. For instance, the scaled standard error from the effect size bootstrap for the variance component from the empty model equals $0.141 * (0.239/0.230) = 0.147$. The example illustrates that the different bootstrap methods can give dissimilar results. To evaluate the performance of the four methods, as compared with the ordinary RIGLS procedure, we performed a simulation study.

A Simulation Study

The effect size measure used in the simulation study was the standardized mean difference, expressing the difference between two groups—for instance, a control and an experimental group. Data were generated in the following way. First, k Level 2 residuals u_j were drawn from a normal distribution with zero mean and variance ^{2} and were added to a mean effect size $= 0.5$, to obtain k true effect sizes $_j$. Next, raw data for both groups in study j were simulated from

$$N \quad \frac{j}{2}, 1$$

and

$$N \ \frac{j}{2}, 1 \ ,$$

respectively, and were summarized by calculating the standardized mean difference and the corresponding standard error, using the formulae of Hedges (1981). To obtain a detailed picture of the performance of the bootstrap methods, different values for the mean group size, the number of studies, and the between-study variance were used (Table 3). Values for the mean population effect size and the between-study variance are comparable

Table 3
Characteristics of the Simulated Data Sets

Mean study group sizes: $\overline{n} = 5 / 25 / 100$
Number of studies: $k = 5 / 10 / 50$
Variance in true effect sizes: $^2 = 0 / 0.05 / 0.1$
Mean population effect size: $= 0.5$
Lack of balance: $n_E = n_C = n \quad U(0.4\,\overline{n}, 1.6\,\overline{n})$
Distribution of true effect sizes and of raw data: normal

with those found in meta-analyses or in other simulation studies. The groups within a study are of equal size, but studies within a data set could differ in size: Group sizes are drawn from a uniform distribution, $U(0.4\overline{n},1.6\overline{n})$, with $\overline{n}$ equal to 5, 25, or 100. Although it is uncommon, in meta-analysis, that the study group size or the number of studies is as small as 5, these situations are included in the simulation design in order to get a better idea of the performance of the bootstrap for small samples.

Three factors, each with three levels, thus contributed to the simulation design: the size of the studies, the number of studies, and the between-study variance. For each combination, 1,000 meta-analytic data sets were generated—in total, 27,000 data sets. Data were generated and analyzed with MLwiN. The bootstrap methods were applied in the same way as that described for the example data.

In order to get stable standard error estimates, we drew at least 300 samples for which the RIGLS algorithm converged. If, after 300 samples, the bootstrap had not yet converged, more samples were drawn. If, after drawing 10,000 bootstrap samples, the bootstrap still had not converged, the sampling was stopped, and the data were discarded. To evaluate convergence of the bootstrap, we kept *running means* of the four parameter estimates (in this case, the overall effect size, the between-study variance, and the corresponding standard error estimates), which, in fact, were the bootstrap estimates of the parameters. Bootstrap estimation was considered as converged if all four bootstrap estimates hardly changed for three successive times when an additional bootstrap sample was drawn:

$$\left| \overline{\hat{\ }}_{t\ i+1} - \overline{\hat{\ }}_{t\ i} \right| < 0.0001, \qquad (11)$$

where $i = 1, 2, 3,$ en and $\overline{\hat{\theta}}_t$ was the mean parameter estimate of the first t bootstrap samples.

We found that the number of (retained) bootstrap samples required for the bootstrap to converge increases with a decreasing n or k and an increasing ^{2}. For the effect size bootstrap, for instance, the average number varied from 1,904 samples for $k = 5$ and $n = 5$ to 208 samples for $k = 50$ and $n = 100$. The number of bootstrap samples required was somewhat larger for the parametric bootstraps (on average, 734 and 775 samples for the effect size and the raw data bootstraps) than for the nonparametric bootstraps (727 and 666 for the error and the cases bootstraps). We note that the number of bootstrap samples that was discarded because of nonconvergence or a negative definite covariance matrix appeared to be very small for all the bootstrap methods (especially for larger k, n and ^{2}). When k and n were 5 and the true variance was zero, about 7% of the bootstrap samples from the cases bootstrap were discarded, because the covariance matrix had gone negative definite, but for the other bootstrap methods (or for other kinds of data), this percentage was usually much smaller and often close to zero.

Results. For large n and k (100 and 50, respectively), the ordinary RIGLS algorithm, as well as the four bootstrap methods, yield unbiased and equally precise parameter estimates and standard errors. The most important differences between the algorithms for smaller n or k can be summarized as follows.

1. As was expected, because of the dependence of the effect size and its sampling variance, the initial RIGLS estimates of the *mean population effect size* were found to be negatively *biased*, especially if n was small (Table 4). Increasing k for a small n even slightly (not shown in the table) exacerbated the problem.

Table 4 furthermore reveals that in the parametric bootstrap methods, the uncorrected bootstrap estimates show an additional negative bias, resulting in relatively unbiased (noniteratively) bias-corrected effect size estimates. Differences between parametric bootstrap methods are small, although the raw data bootstrap, in which the underlying sampling process is more closely simulated, performs somewhat better in cases in which the bias is most severe, when n is small and k large (not visible in the table). Also, the error bootstrap succeeds in eliminating most of the bias. In the cases bootstrap, however, the bootstrap sample estimates show very little additional bias, and the bias correction thus has almost no effect.

2. There is a tradeoff between bias and *precision* when the mean effect size is estimated: The variation of the estimates is smallest for the highly biased uncorrected parametric bootstrap estimates and largest for the unbiased corrected parametric bootstrap estimates. Whether correction for bias is recommended depends primarily on the number of studies, since with an increasing number of studies, bias increases slightly and precision decreases. This can be seen in Table 5, presenting the MS_e of the estimates for meta-analyses of small studies ($n = 5$). If the number of studies is small ($k = 5$ or 10), the MS_e of the uncorrected parametric bootstrap estimates is generally smaller than that of the initial RIGLS

Table 4
Mean of the Overall Effect Size Estimates ($\delta = 0.5$)

	Initial	Uncorrected				Bias Corrected			
n	RIGLS	ES	RD	Error	Cases	ES	RD	Error	Cases
5	0.455	0.411	0.409	0.409	0.454	0.499	0.502	0.501	0.456
25	0.494	0.484	0.485	0.484	0.494	0.504	0.503	0.504	0.494
100	0.499	0.496	0.496	0.496	0.499	0.501	0.501	0.501	0.499

Note—Estimates are averaged over k and ^{2}. ES, effect size bootstrap; RD, raw data bootstrap; Error, error bootstrap; Cases, cases bootstrap.

Table 5
MS_e *10,000 of the Mean Effect Size Estimates (for $n = 5$)

	Initial	Uncorrected				Bias Corrected			
k	RIGLS	ES	RD	Error	Cases	ES	RD	Error	Cases
5	844	792	750	790	882	998	988	1,009	850
10	440	426	418	428	446	520	519	529	451
50	108	162	171	166	113	96	98	98	104

Note—ES, effect size bootstrap; RD, raw data bootstrap; Error, error bootstrap; Cases, cases bootstrap.

estimates, whereas that of the corrected estimates is larger. If the number of studies is relatively large (50), the MS_e of the uncorrected parametric bootstrap estimates is generally larger, whereas that of the corrected estimates is smaller. The same is true for the error bootstrap. The MS_e for both the corrected and the uncorrected cases bootstrap estimates is usually higher than that for the ordinary RIGLS estimation procedure. For larger n, the findings are similar, although less pronounced.

3. We found negative bias when using the RIGLS estimation procedure to estimate the *between-study variance*, as has been observed before (Busing, 1993; Mok, 1995; Raudenbush & Bryk, 2002). Table 6, presenting the mean (untruncated) estimates in the case in which the population variance is 0.1, shows that the bias is largest when there are many small studies. The bias correction from the error bootstrap, the raw data bootstrap, and especially, the effect size bootstrap eliminates a large part of the bias. In the cases bootstrap, the uncorrected bootstrap estimates show additional negative bias due to ties in the data. Since the influence of ties is more pronounced for small k, the bias is overestimated when k is small and underestimated when k is large, resulting in biased corrected estimates.

4. As with the mean effect size estimates, there is a tradeoff between bias and precision for the *between-study variance* estimates: The variation of the initial RIGLS estimates is larger than that of the uncorrected bootstrap estimates but smaller than that of the bias-corrected estimates. In Table 7, the mean squared error of the truncated bootstrap estimates is given.

The uncorrected variance estimates of all the bootstrap methods show a decreased mean squared error, as compared with the initial RIGLS estimates, especially when k and n are small. As compared with the other bootstrap methods, the uncorrected cases bootstrap estimates are somewhat better for small k and n. For all the methods, the correction for bias results in an increase of the MS_e. The bias-corrected estimates show a larger MS_e than the initial RIGLS estimates do.

Table 6
Mean of the (Untruncated) Between-Study Variance Estimates ($^2 = 0.1$)

		Initial	Uncorrected				Bias Corrected			
n	k	RIGLS	ES	RD	Error	Cases	ES	RD	Error	Cases
5	5	0.032	0.074	0.097	0.086	−0.030	0.094	0.071	0.083	0.198
	10	0.020	0.018	0.022	0.024	−0.001	0.094	0.090	0.088	0.114
	50	0.004	−0.037	−0.041	−0.040	−0.001	0.078	0.082	0.081	0.042
25	5	0.095	0.101	0.102	0.101	0.061	0.099	0.097	0.098	0.139
	10	0.096	0.092	0.094	0.093	0.079	0.102	0.101	0.101	0.115
	50	0.093	0.086	0.086	0.086	0.090	0.101	0.100	0.101	0.097
100	5	0.096	0.096	0.096	0.096	0.073	0.096	0.096	0.096	0.120
	10	0.098	0.097	0.097	0.098	0.086	0.099	0.099	0.099	0.110
	50	0.099	0.098	0.098	0.098	0.097	0.101	0.101	0.101	0.102

Note—ES, effect size bootstrap; RD, raw data bootstrap; Error, error bootstrap; Cases, cases bootstrap.

Table 7
MS_e *10000 of the (Truncated) Between-Study Variance Estimates

		Initial	Uncorrected				Bias Corrected			
n	k	RIGLS	ES	RD	Error	Cases	ES	RD	Error	Cases
5	5	634	408	454	428	388	927	855	898	1037
	10	235	117	115	123	200	426	430	415	298
	50	35	38	38	38	35	53	56	54	33
25	5	86	81	82	82	53	92	90	91	139
	10	40	36	37	37	33	44	43	44	51
	50	7	8	8	8	7	8	8	8	8
100	5	33	33	33	33	24	34	33	33	50
	10	16	16	16	16	14	16	16	16	20
	50	3	3	3	3	3	3	3	3	3

Note—ES, effect size bootstrap; RD, raw data bootstrap; Error, error bootstrap; Cases, cases bootstrap.

Table 8
Mean of the Standard Error Estimates of the Mean Effect Size

			Initial	Initial	Unscaled				Scaled			
^{2}	*n*	*k*	RIGLS *SD*	RIGLS *SE*	ES	RD	Error	Cases	ES	RD	Error	Cases
0	5	5	0.271	0.317	0.296	0.304	0.303	0.249	0.457	0.365	1.235	0.300
		50	0.084	0.094	0.084	0.084	0.084	0.084	0.104	0.106	0.105	0.086
	100	5	0.064	0.073	0.074	0.074	0.074	0.060	0.075	0.075	0.075	0.060
		50	0.021	0.021	0.021	0.021	0.021	0.020	0.021	0.021	0.021	0.020
0.1	5	5	0.307	0.333	0.311	0.318	0.317	0.279	0.457	0.402	0.498	0.426
		50	0.096	0.097	0.087	0.087	0.087	0.093	0.106	0.110	0.107	0.094
	100	5	0.155	0.145	0.145	0.145	0.145	0.131	0.147	0.147	0.147	0.131
		50	0.050	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050	0.049

Note—The high value for the mean scaled standard error for the error bootstrap is partly due to one outlier. Without the outlier, the value becomes 0.598. ES, effect size bootstrap; RD, raw data bootstrap; Error, error bootstrap; Cases, cases bootstrap.

5. The mean of the *standard errors of the mean effect size* are presented in Table 8 for some combinations of *n*, *k* and ^{2}. Because the standard errors represent the standard deviations of the sampling distributions of the estimates (Efron, 1982), we can evaluate the estimated standard errors by comparing them with the standard deviations of the initial RIGLS estimates, which are also presented in Table 8. We note that the initial RIGLS standard errors are accurate only if both *k* and *n* are large. Standard errors are positively biased if *n* is small or when the true variance is zero, especially if *k* is small. Standard errors are negatively biased otherwise.

The unscaled bootstrap standard errors are substantially lower than the initial RIGLS standard errors, particularly when the studies are small. This is especially the case for the nonparametric cases bootstrap standard errors, probably due to ties. Table 8 shows that scaling the standard errors has a small effect when *n* is large (because in this situation, the estimate of the mean effect size shows small bias and, therefore, the scaling factor is close to 1), whereas for small *n*, standard errors are generally inflated when scaled. The unscaled standard errors of the parametric bootstrap methods and of the error bootstrap are, in general, closer to the standard deviations of the initial estimates than are the ordinary standard error estimates if *n* is 5, although they are still somewhat too large, but the scaled standard errors are even more positively biased than the initial standard error estimates are. Again, the cases bootstrap performs somewhat worse than the other procedures.

Finally, we note that for small *n*, the unscaled standard errors of all the bootstrap methods vary slightly less than the RIGLS standard errors, whereas the scaled standard errors show a much larger variation.

6. To evaluate the *standard error estimates for the between-study variance*, we compared them with the standard deviation of the (untruncated) variance estimates. We found that the standard errors given by the ordinary RIGLS algorithm are in general too high, especially when the true variance is small and *n* and *k* are small. The unscaled standard errors obtained by the parametric bootstrap methods amount to a substantial improvement, although they are often still somewhat too high. The error bootstrap gives standard errors that are relatively accurate, although they are generally somewhat too small. The uncorrected standard errors from the cases bootstrap are generally too small and do not amount to an improvement, as compared with those of the ordinary RIGLS procedure. Finally, the scaled standard errors of all the bootstrap methods are generally much too high, unless the variance, *n,* and *k* are large.

Discussion and Conclusions

In traditional meta-analyses, *fixed* effects models are often used, assuming that the effect one is interested in is exactly the same in each study or that all heterogeneity is accounted for by the study characteristics included in the model. These assumptions are often unrealistic, resulting in biased standard errors and, possibly, in wrong conclusions. Alternatively, one could assume a *random* effects model, regarding study effects as a random sample from a population of effects. In this model, study effects are assumed to be exchangeable, which again is a strong assumption, since study characteristics are often likely to moderate the effect. A *mixed* effects model or a multilevel model (Equation 2) combines the strengths of both kinds of models and frees the restrictive assumptions corresponding to the fixed and the random effects models (National Research Council, 1992): Effects are described by using study characteristics, but at the same time, a possible residual between-study variance is explicitly modeled. The multilevel model can easily be extended further—for instance, by including additional levels. As a result, more complex questions that cannot be managed by traditional approaches can be handled in the multilevel approach. The parameters of multilevel models usually are estimated using maximum likelihood. The purpose of this article has been to show that adaptations of bootstrap methods developed for multilevel models can be used in meta-analysis, in order to overcome some undesirable characteristics of the maximum likelihood estimates. We have proposed two parametric and two nonparametric bootstrap methods that can be used for meta-analysis.

In a simulation study, we found that only if the number and the size of studies are large (50 studies, with av-

erage group sizes of 100 or more) does the ordinary restricted maximum likelihood give unbiased estimates and correct standard errors. For smaller data sets, (restricted) maximum likelihood often results in biased parameter estimates and standard errors, although it has been suggested before (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) that maximum likelihood can validly be used for combining the results of 20 or 30 studies with group sizes of 15. We want to note that maximum likelihood shares the poorer performance for relatively small meta-analytic data sets with traditional estimation procedures for random effects models. Van den Noortgate and Onghena (2003b) compared the use of multilevel models without study characteristics with the random effects approaches proposed by Hedges and Olkin (1985) and by DerSimonian and Laird (1986) and found that the results were very similar. Approaches for fixed effects models are still valid when there is a relatively small number of studies (Van den Noortgate & Onghena, 2003b), but the use of fixed effects models is restricted to very specific situations, as was discussed above.

The results of the parametric bootstrap methods and the error bootstrap are very similar. For large data sets, they give the same results as the maximum likelihood method, whereas for smaller data sets, these methods eliminate at least a large part of the bias in estimating the mean effect size, which occurs (for small studies) if the observed effect sizes and the corresponding estimated standard errors are dependent. Also, the negative bias in estimating the between-study variance (especially for a large number of small studies and a large between-study variance) is largely eliminated. An interesting result of our simulation study is that correcting the bootstrap estimates for bias decreases precision, possibly resulting in an increased MS_e, a problem that is discussed in Efron and Tishirani (1993). Therefore, although more biased, uncorrected bootstrap estimates of the between-study variance are to be preferred over ordinary RIGLS estimates and over bias-corrected bootstrap estimates. For estimating mean effect size, bias-corrected bootstrap estimates reduce the MS_e only if the number of studies is large and, thus, are to be preferred in this situation. If the number of studies is small or moderate, the uncorrected bootstrap is to be preferred. The parametric bootstrap and the error bootstrap, furthermore, improve the standard error estimates for both the mean effect size and the between-study variance from the ordinary maximum likelihood procedure. Scaling the standard errors does not seem to be recommended. Differences between both parametric methods are very small, although the raw data bootstrap succeeds somewhat better in eliminating the bias in estimating the mean effect size. Note that using an effect size measure with a sampling distribution that is unknown or less well approximated by a normal distribution may make the strength of the raw data bootstrap more pronounced, as compared with the effect size bootstrap. As Efron (1982) has said, "The charm of the jackknife and the bootstrap is that they can be applied to complicated situations where parametric modeling and/or theoretical analysis is hopeless" (p. 28).

The performance of the nonparametric cases bootstrap was found disappointing, except for combining a large number of large studies. The cases bootstrap fails to remove any bias in the mean effect size estimates. Moreover, the cases bootstrap does not yield accurate estimates of (and corrections for) the bias of the ordinary between-study variance estimate. Estimated standard errors for the mean effect size and the between-study variance are often much too small, a conclusion that is in line with the statement of Laird and Louis (1987) that the nonparametric bootstrap may be too liberal. The poor performance of the cases bootstrap may be partly due to the problem that, for meta-analysis, only cases on the second level can be drawn.

Looking back at the example, we see that also for the open-education data from Hedges and Olkin (1985), the results of the cases bootstrap differ from those of the other estimation procedures. The bootstrap results further suggest that fixed coefficients and the between-study variance estimates from the RIGLS procedure are biased and that the true values are somewhat further from zero. The results of our simulation study imply that for this kind of data set, consisting of a small set of studies of moderate size, parameter estimates should not be corrected for bias, and the uncorrected parametric or the error bootstrap parameter estimates are to be preferred.

In general, the simulation study puts forward the superiority of the parametric bootstrap and the nonparametric error bootstrap over the ordinary RIGLS procedure, unless the results of a large set of large studies are to be combined. Even for very small meta-analytic data sets, consisting of five studies in which two groups of 5 study participants are compared, the bootstrap procedures seem to perform relatively well. A limitation of our simulation study is, however, that raw data were simulated from normal distributions, fulfilling an assumption of the RIGLS estimation and the parametric bootstrap methods, and the results are, in principle, restricted to this condition. More research is needed to evaluate the performance of the bootstrap methods when the normality assumption is violated.

REFERENCES

Busing, F. M. T. A. (1993). *Distribution characteristics of variance estimates in two-level models.* Leiden: Psychometrics and Research Methodology.

Busing, F. M. T. A., Meijer, E., & van der Leeden, R. (1994). *MLA: Software for multilevel analysis of data with two levels. User's guide for Version 1.0b*. Leiden: Leiden University.

Carpenter, J., Goldstein, H., & Rasbash, J. (1999). A non-parametric bootstrap for multilevel models. *Multilevel Modelling Newsletter*, **11**, 2-5.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, **112**, 155-159.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application.* Cambridge: Cambridge University Press.

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

EFRON, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: SIAM.

EFRON, B., & TIBSHIRANI, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

GOLDSTEIN, H. (1995). *Multilevel statistical models*. London: Edward Arnold.

GOLDSTEIN, H. (1996). Consistent estimators for multilevel generalized linear models using an iterated bootstrap. *Multilevel Modelling Newsletter*, **8**, 3-6.

GOLDSTEIN, H., YANG, M., OMAR, R., TURNER, R., & THOMPSON, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society: Series C*, **49**, 339-412.

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). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics*, **7**, 119-137.

HEDGES, L. V., GIACONIA, R. M., & GAGE, N. L. (1981). *The empirical evidence on the effectiveness of open education*. Stanford, CA: Stanford University School of Education.

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

KUK, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society: Series B*, **57**, 395-407.

LAIRD, N. M., & LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, **82**, 739-759.

LAIRD, N. M., & LOUIS, T. A. (1989). Empirical Bayes confidence intervals for a series of related experiments. *Biometrics*, **45**, 481-495.

MEIJER, E., VAN DER LEEDEN, R., & BUSING, F. M. T. A. (1995). Implementing the bootstrap for multilevel models. *Multilevel Modelling Newsletter*, **7**, 7-11.

MOK, M. (1995). Sample size requirements for 2-level designs in educational research. *Multilevel Modelling Newsletter*, **7**, 11-15.

NATIONAL RESEARCH COUNCIL (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy Press.

RASBASH, J., BROWNE, W., GOLDSTEIN, H., YANG, M., PLEWIS, I., HEALY, M., WOODHOUSE, G., DRAPER, D., LANGFORD, I., & LEWIS, T. (2000). *A user's guide to MLwiN*. London: University of London.

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

RAUDENBUSH, S. W., & BRYK, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

SMITH, S. J., CAUDILL, S. P., STEINBERG, K. S., & THACKER, S. B. (1995). On combining dose-response data from epidemiological studies by meta-analysis. *Statistics in Medicine*, **14**, 531-544.

SNIJDERS, T. A. B., & BOSKER, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

VAN DEN NOORTGATE, W., & ONGHENA, P. (2003a). Estimating the standardized mean difference in a meta-analysis: Bias, precision and mean squared error. *Behavior Research Methods, Instruments, & Computers*, **35**, 504-511.

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

VAN DEN NOORTGATE, W., & ONGHENA, P. (2003c). A parametric bootstrap version of Hedges' homogeneity test. *Journal of Modern Applied Statistical Methods*, **2**, 73-79.

ZHOU, X.-H., BRIZENDINE, E. J., & PRITZ, M. B. (1999). Methods for combining rates from several studies. *Statistics in Medicine*, **18**, 557-566.

NOTES

1. According to Cohen (1992), effect sizes of 0.20, 0.50, and 0.80 refer to small, moderate and large effects, respectively.

2. Since the sampling distribution of variance estimates is only very approximately normal (as was discussed above), using the Wald is not recommended for testing variance components. Instead, one could use the likelihood ratio test, which shows, for the example, that the between-study heterogeneity is statistically highly significant [$^2(1) = 25.20$, $p < .001$].

(Manuscript received January 14, 2003;
revision accepted for publication June 27, 2004.3)