Received 12 March 2010, Revised 1 July 2010, Accepted 29 July 2010 Published online 4 October 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.11

Outlier and influence diagnostics for meta-analysis

Wolfgang Viechtbauer[a*†] and Mike W.-L. Cheung[b]

The presence of outliers and influential cases may affect the validity and robustness of the conclusions from a meta-analysis. While researchers generally agree that it is necessary to examine outlier and influential case diagnostics when conducting a meta-analysis, limited studies have addressed how to obtain such diagnostic measures in the context of a meta-analysis. The present paper extends standard diagnostic procedures developed for linear regression analyses to the meta-analytic fixed- and random/mixed-effects models. Three examples are used to illustrate the usefulness of these procedures in various research settings. Issues related to these diagnostic procedures in meta-analysis are also discussed. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: meta-analysis; outliers; influence diagnostics; mixed-effects model

1. Introduction

Meta-analysis is the statistical analysis of effect sizes obtained from a pool of empirical studies. The goal of aggregating the results from related studies is to obtain information about the overall effect and to examine the influence of study-level characteristics on the size of the effect. The underlying statistical methods for conducting meta-analyses are generally well-developed (see, e.g. [1]) and applied on a regular basis in a variety of different disciplines (e.g. psychology, medicine, epidemiology, ecology, business/consumer research).

Similar to other types of data, it is not uncommon to observe extreme effect size values when conducting a meta-analysis. As the main objective of a meta-analysis is to provide a reasonable summary of the effect sizes of a body of empirical studies, the presence of such outliers may distort the conclusions of a meta-analysis. Moreover, if the conclusions of a meta-analysis hinge on the data of only one or two influential studies, then the robustness of the conclusions are called into question.

Researchers, therefore, generally agree that the effect sizes should be examined for potential outliers and influential cases when conducting a meta-analysis [2–5]. The most thorough treatment of outlier diagnostics in the context of meta-analysis to date can be found in the classic book by Hedges and Olkin [2], who devoted a whole chapter to diagnostic procedures for effect size data. Several graphical methods have also been proposed to inspect the data for unusual cases (e.g. [6, 7]). However, the methods developed by Hedges and Olkin [2] are only applicable to fixed-effects models. Given that random- and mixed-effects models are gaining popularity in the meta-analytic context, corresponding methods for outlier and influential case diagnostics need to be developed.

The present paper introduces several outlier and influence diagnostic procedures for the random- and mixed-effects model in meta-analysis. These procedures are logical extensions of the standard outlier and case diagnostics for regular regression models and take both sampling variability and between-study heterogeneity into account. The proposed measures provide a simple framework for evaluating the potential impact of outliers or influential cases in meta-analysis.

The paper is organized as follows. In the next section, we provide a brief review of various meta-analytic models, including the random- and mixed-effects model. In Sections 3 and 4, we then show how conventional diagnostic procedures for outlier and influential case detection in standard linear regression can be extended to these models. Three examples are then used to illustrate how to apply these procedures in practice. Finally, the discussion section touches upon some important issues relating to these diagnostic methods and provides some directions for future research.

[a] *Department of Methodology and Statistics, Maastricht University, Maastricht, The Netherlands*

[b] *Department of Psychology, National University of Singapore, Singapore, Singapore*

[*] *Correspondence to: Wolfgang Viechtbauer, Department of Methodology and Statistics, Maastricht University, Maastricht, The Netherlands.*

[†] *E-mail: wolfgang.viechtbauer@stat.unimaas.nl*

2. Meta-analytic models

In this section, we review various meta-analytic models, discuss how these models can be fitted, and show how to obtain predicted (average) effects under these models.

2.1. Description of the models

Let $y_1,\ldots,y_k$ denote the observed effect size estimates in a set of k independent studies. We use the term 'effect size' generically here, so the y_i values may consist of a set of (standardized) mean differences, raw correlation coefficients or their Fisher z-transformed counterparts, (log) odds or risk ratios, or any other outcome measure typically employed in meta-analyses [8, 9].

We will assume that

$$y_i = \theta_i + e_i, \tag{1}$$

where $e_i \sim N(0, v_i)$ and θ_i denotes the true effect in the ith study. Therefore, the observed effect size in the ith study is assumed to be an unbiased and normally distributed estimate of the corresponding true effect with sampling variance equal to v_i. The sampling variances are assumed to be known. Depending on the outcome measure chosen, one must rely on the asymptotic behavior of the estimator for these assumptions to be approximately justified. For certain outcome measures, it may also be helpful to first apply a bias correction, variance stabilizing, and/or normalizing transformation to ensure that the assumptions are approximately satisfied. For example, Hedges [10] demonstrated how the standardized mean difference can be corrected for its slight positive bias, while Fisher's r-to-z-transformation [11] is a variance stabilizing and normalizing transformation for correlation coefficients.

The true effects may be homogeneous (i.e. $\theta_i = \theta$ for all i) or heterogeneous (i.e. not all θ_i equal to each other). Cochran [12] proposed the so-called Q statistic to test the homogeneity of the effect sizes, which is given by

$$Q = \sum w_i (y_i - \hat{\theta})^2, \tag{2}$$

where $w_i = 1/v_i$ and $\hat{\theta} = \sum w_i y_i / \sum w_i$ is the inverse-variance weighted estimate of θ under the assumption of homogeneity (all summations go from $i=1$ to k throughout the paper unless otherwise noted). Under the null hypothesis that all effect sizes are homogeneous, the Q statistic follows a chi-square distribution with $k-1$ degrees of freedom.

Homogeneity may be a reasonable assumption when the studies to be meta-analyzed are (near) replicates of each other and were conducted with samples coming from similar populations. However, this will often not be the case and differences in the methods and the characteristics of the samples may introduce heterogeneity into the true effects [13, 14]. One possibility is to consider the heterogeneity to be a result of purely random processes.

This approach leads to the random-effects model, where we assume that

$$\theta_i = \mu + u_i \tag{3}$$

and $u_i \sim N(0, \tau^2)$, so that μ denotes the average true effect and τ^2 denotes the amount of heterogeneity in the true effects [15]. Assuming independence between e_i and u_i, it follows that $y_i \sim N(\mu, \tau^2 + v_i)$. The goal is then to estimate μ and τ^2. Note that $\tau^2 = 0$ implies homogeneity among the true effects, so that $\mu \equiv \theta$.

If the variables moderating the size of the effect are known, then we can model the relationship between the effect sizes and moderators. Typically, we then assume that the true effect for a particular study is a (linear) function of a set of moderators plus a certain amount of residual heterogeneity, so that

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + u_i, \tag{4}$$

where x_{ij} denotes the value of the jth moderator variable for the ith study, β_0 denotes the expected effect size when $x_{ij} = 0$ for $j = 1,\ldots,p$, and β_j denotes how the jth moderator influences the size of the true effect for a one-unit increase in x_{ij}. We still assume that $u_i \sim N(0, \tau^2)$, but τ^2 should now be interpreted as the amount of residual heterogeneity in the effect sizes (that is, the amount of variability among the true effects not accounted for by the influence of the moderators included in the model). Since the true effect sizes are now considered to be a function of both fixed and random effects, this model is typically called the mixed-effects model in the meta-analytic literature [16]. Analyses employing such models are typically called meta-regression analyses [17, 18].

Note that the mixed-effects model simplifies to the random-effects model when $\beta_1 = \cdots = \beta_p = 0$. Also, $\tau^2 = 0$ now implies that all of the heterogeneity among the true effects is a result of the moderators included in the model.

2.2. Fitting the meta-analytic models

To fit the random-effects model, we first estimate τ^2 with one of the various estimators that have been developed for this purpose (see, e.g. [19] for a comparison of various heterogeneity estimators). One of the most commonly used estimators of τ^2 was proposed by DerSimonian and Laird [20]. The estimator is given by

$$\hat{\tau}^2 = \frac{Q - (k-1)}{c}, \tag{5}$$

where Q is the statistic of the homogeneity test given in (2) and $c = \sum w_i - \sum w_i^2 / \sum w_i$. When the estimated value is negative, it is truncated to zero.

Once the amount of heterogeneity is estimated, μ can be estimated with

$$\hat{\mu} = \frac{\sum \tilde{w}_i y_i}{\sum \tilde{w}_i}, \tag{6}$$

where the weights are now given by $\tilde{w}_i = 1/(v_i + \hat{\tau}^2)$. The variance of $\hat{\mu}$ is approximately equal to

$$\text{Var}[\hat{\mu}] = \frac{1}{\sum \tilde{w}_i}. \tag{7}$$

An approximate 95% confidence interval for μ can be obtained with

$$\hat{\mu} \pm 1.96\sqrt{\text{Var}[\hat{\mu}]}. \tag{8}$$

See [21] for an adjustment to this approach that provides a confidence interval for μ with slightly better coverage probability, especially when k is small [22].

DerSimonian and Laird's estimator can be generalized to the meta-analytic mixed-effects model [16]. Using matrix notation considerably simplifies the notation. Let $\boldsymbol{y}$ denote the column vector of the effect size estimates for the k studies. Next, we define $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]'$ and $\boldsymbol{X}$ as the $k \times (p+1)$ design matrix that includes 1's in the first column (corresponding to β_0) and the values of the moderators in the other p columns (corresponding to β_1 through β_p). Finally, let $\boldsymbol{V} = \text{diag}[v_1, v_2, \ldots, v_k]$ denote a $k \times k$ diagonal matrix with the sampling variances of the effect size estimates.

Now let $Q_E = \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}$, where $\boldsymbol{P} = \boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}$ and $\boldsymbol{W} = \boldsymbol{V}^{-1}$. The amount of residual heterogeneity can then be estimated with

$$\hat{\tau}^2 = \frac{Q_E - (k-p-1)}{\text{trace}[\boldsymbol{P}]}, \tag{9}$$

with negative values of $\hat{\tau}^2$ again truncated to zero. Once the amount of residual heterogeneity is estimated, we can estimate the vector of regression coefficients via weighted least squares using

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\tilde{\boldsymbol{W}}X)^{-1}\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{y}, \tag{10}$$

where $\tilde{\boldsymbol{W}} = \text{diag}[1/(v_1 + \hat{\tau}^2), 1/(v_2 + \hat{\tau}^2), \ldots, 1/(v_k + \hat{\tau}^2)]$. The variance–covariance matrix of the parameter estimates in $\hat{\boldsymbol{\beta}}$ can be estimated with

$$\text{Var}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X})^{-1}. \tag{11}$$

Approximate 95% confidence intervals for the model coefficients can be obtained with

$$\hat{\beta}_j \pm 1.96\sqrt{\text{Var}[\hat{\beta}_j]}, \tag{12}$$

where $\text{Var}[\hat{\beta}_j]$ is the corresponding diagonal element from $\text{Var}[\hat{\boldsymbol{\beta}}]$ (see [23] for a slight adjustment to this approach, which provides slightly better coverage probabilities of the confidence intervals). The null hypothesis $H_0: \tau^2 = 0$ under the mixed-effects model can be tested by comparing the Q_E statistic against the critical value of a chi-square distribution with $k-p-1$ degrees of freedom.

Note that the random-effects model is actually just a special case of the mixed-effects model and is obtained by setting $\boldsymbol{X}$ equal to a column vector of 1's (Q_E then simplifies to Q, $\hat{\tau}^2$ given by (9) then simplifies to the DerSimonian and Laird estimator given by (5), and $\hat{\boldsymbol{\beta}}$ simplifies to $\hat{\mu}$).

2.3. Predicted (average) effects

Under the mixed-effects model, $E[\theta_i | x_{i1}, \ldots, x_{ip}] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$, which we will simply denote as μ_i. Therefore, having obtained estimates of the regression coefficients, we can estimate the average true effect for the ith study with $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$. Alternatively, the predicted average true effects may also be obtained from the observed effect sizes via the hat matrix. In particular, for the entire set of k studies, we can write this in matrix notation as $\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{y}$, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\tilde{\boldsymbol{W}}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{W}}$ is the hat matrix. The variance of $\hat{\mu}_i$ can be estimated with

$$\text{Var}[\hat{\mu}_i] = h_i(v_i + \hat{\tau}^2), \tag{13}$$

where h_i is the ith diagonal element of $\boldsymbol{H}$ (the so-called 'leverage' of the ith study). When $\hat{\tau}^2 = 0$ under the mixed-effects model, then this suggest that all of the heterogeneity can be accounted for by the moderators included in the model, so $\hat{\mu}_i \equiv \hat{\theta}_i$ then denotes the estimated true effect (as opposed to the estimated *average* true effect) for a particular combination of moderator values.

In the random-effects model, the predicted average true effect is the same for all k studies, namely $\hat{\mu}_i = \hat{\mu}$, with variance equal to $\text{Var}[\hat{\mu}]$ given by (7). The diagonal elements of the hat matrix are, therefore, equal to $h_i = \tilde{w}_i / \sum \tilde{w}_i$. Here, the meaning of the term 'leverage' becomes quite apparent, as the study with the largest h_i value is also the study that exerts the largest influence on $\hat{\mu}$.

Finally, note that $\hat{\tau}^2 = 0$ under a random-effects model suggests homogeneity of the true effects, so that $\hat{\mu}_i = \hat{\theta}$ then denotes the predicted true effect for all k studies. The leverages are then equal to $h_i = w_i / \sum w_i$.

3. Identifying outliers in a meta-analysis

Many meta-analyses will include at least a few studies yielding observed effects that appear to be outlying or extreme in the sense of being well separated from the rest of the data. Visual inspection of the data may be one way of identifying unusual cases, but this approach may be problematic especially when dealing with models involving one or more moderators. Moreover, the studies included in a meta-analysis are typically of varying sizes (and hence, the sampling variances of the y_i values will differ), further complicating the issue. A more formal approach for identifying outliers is based on an examination of the residuals in relation to their corresponding standard errors.

Various types of residuals have been defined in the context of linear regression [24], which can be adapted to the meta-analytic random- and mixed-effects model (and any special cases thereof). One residual to consider is the (internally) studentized residual, given by

$$s_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{Var}[y_i - \hat{\mu}_i]}}, \tag{14}$$

where y_i and $\hat{\mu}_i$ are the observed and the predicted (average) effect size for the ith study, $e_i = y_i - \hat{\mu}_i$ is the raw residual, and $\text{Var}[y_i - \hat{\mu}_i]$ is the sampling variance of the raw residual, which is equal to $\text{Var}[y_i - \hat{\mu}_i] = (1 - h_i)(v_i + \hat{\tau}^2)$. Note that $\hat{\mu}_i$ is simply equal to $\hat{\mu}$ for all k studies in the random-effects model.

As y_i is involved in the calculation of $\hat{\mu}_i$, it may have a large influence on $\hat{\mu}_i$ especially if y_i deviates strongly from the assumed model. In fact, if the ith study is indeed an outlier, then $\hat{\mu}_i$ will be pulled toward the y_i value, making it more difficult to identify the outlying study. In addition, the presence of an outlier will lead to an inflated estimate of τ^2, which in turn results in an overestimation of the sampling variance of the residual (and hence, a studentized residual that is too small), making it again more difficult to detect the outlier.

Consequently, following the suggestion of Hedges and Olkin for fixed-effects models [2], we recommend to use the studentized deleted (or also called the externally studentized) residuals, given by

$$t_i = \frac{y_i - \hat{\mu}_{i(-i)}}{\sqrt{\text{Var}[y_i - \hat{\mu}_{i(-i)}]}}, \tag{15}$$

where $\hat{\mu}_{i(-i)}$ is the predicted average true effect size for the ith study based on the model that actually excludes the ith study during the model fitting. Therefore, $e_{i(-i)} = y_i - \hat{\mu}_{i(-i)}$ is the so-called 'deleted residual' for the ith case. As y_i and $\hat{\mu}_{i(-i)}$ are uncorrelated, (15) simplifies to

$$t_i = \frac{y_i - \hat{\mu}_{i(-i)}}{\sqrt{v_i + \hat{\tau}^2_{(-i)} + \text{Var}[\hat{\mu}_{i(-i)}]}}, \tag{16}$$

where $\hat{\tau}^2_{(-i)}$ denotes the estimated amount of (residual) heterogeneity and $\text{Var}[\hat{\mu}_{i(-i)}]$ the estimated amount of variability in $\hat{\mu}_{i(-i)}$ from the model that excludes the ith study. For the random-effects model, $\hat{\mu}_{i(-i)}$ should be replaced with $\hat{\mu}_{(-i)}$.

The equation for the studentized deleted residual given by (16) clearly illustrates the three sources of variability that contribute to the difference between the observed effect size y_i and the predicted average true effect when the ith study actually fits the assumed model, namely sampling variability, (residual) heterogeneity among the true effects, and imprecision in the predicted average effect.

If the studies actually follow the assumed model, then the studentized deleted residuals from the set of studies approximately follow a standard normal distribution. On the other hand, a study that does not fit the assumed model will tend to yield an observed effect that deviates more strongly from $\hat{\mu}_{i(-i)}$ (or $\hat{\mu}_{(-i)}$) than would be expected based on these three sources of variability. Hence, its studentized deleted residual will tend to be large.

In fact, just as in standard linear regression (e.g. [25]), the studentized deleted residual for a particular study formalizes a proper outlier test under a mean shift outlier model. In particular, suppose that the true model is given by

$$E[\theta_i | x_{i1}, \ldots, x_{ip}] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{17}$$

except for one study, denoted by $\tilde{i}$, whose model is given by

$$E[\theta_{\tilde{i}} | x_{\tilde{i}1}, \ldots, x_{\tilde{i}p}] = \beta_0 + \beta_1 x_{\tilde{i}1} + \cdots + \beta_p x_{\tilde{i}p} + \delta, \tag{18}$$

where δ denotes a fixed amount by which the expected value of $\theta_{\tilde{i}}$ is shifted away from the true model. A test of $H_0 : \delta = 0$ can be easily obtained by adding a dummy variable to the model (i.e. adding a column to $\boldsymbol{X}$) that is equal to 1 for $i = \tilde{i}$ and 0 otherwise. The parameter estimate corresponding to the dummy variable is then equal to $\hat{\delta} = y_{\tilde{i}} - \hat{\mu}_{\tilde{i}(-\tilde{i})}$, the deleted residual for study $\tilde{i}$, with standard error equal to $SE[\hat{\delta}] = \sqrt{v_{\tilde{i}} + \hat{\tau}^2_{(-\tilde{i})} + \text{Var}[\hat{\mu}_{\tilde{i}(-\tilde{i})}]}$. The studentized deleted residual, therefore, corresponds to the test statistic for the test $H_0 : \delta = 0$. Accordingly, studies with absolute studentized deleted residuals larger than 1.96 may call for a closer inspection.

Naturally, a certain number of studentized deleted residuals are expected to be this large by chance alone. In fact, assuming that the model is correctly specified and no outliers are present in the data, approximately 5% of the studentized deleted residuals would be expected to exceed the bounds ±1.96. Therefore, for example, it would not be surprising to find at least one such value for $k = 20$. However, finding more than two values this large in a set of 20 studies would then only occur in approximately 8% of all meta-analyses (i.e. $\text{Prob}(X > 2) \approx 0.08$, where X follows a binomial distribution with $\pi = 0.05$ and $N = 20$). Based on similar considerations for other values of k, one could consider finding more than $k/10$ studentized deleted residuals larger than ±1.96 in a set of k studies as unusual. However, regardless of the actual number observed, each study with a large studentized deleted residual should be carefully scrutinized. We will return to this point in more detail in the discussion section of this paper.

4. Identifying influential cases in a meta-analysis

An outlying case may not be of much consequence if it exerts little influence on the results. However, if the exclusion of a study from the analysis leads to considerable changes in the fitted model, then the study may be considered to be influential. Case deletion diagnostics known from linear regression (e.g. [24, 26]) can also be adapted to the context of meta-analysis to identify such studies.

Following Belsley *et al.* [26], we can examine the difference between the predicted average effect for the ith study once with and once without the ith study included in the model fitting. Dividing this difference by the standard error of $\hat{\mu}_i$, but replacing $\hat{\tau}^2$ with $\hat{\tau}^2_{(-i)}$, yields

$$\text{DFFITS}_i = \frac{\hat{\mu}_i - \hat{\mu}_{i(-i)}}{\sqrt{h_i(v_i + \hat{\tau}^2_{(-i)})}}, \tag{19}$$

which essentially indicates by how many standard deviations the predicted average effect for the ith study changes after excluding the ith study from the model fitting. For random-effects models, we simply need to replace $\hat{\mu}_{i(-i)}$ with $\hat{\mu}_{(-i)}$ and $\hat{\mu}_i$ with $\hat{\mu}$.

To examine what effect the deletion of the ith study has on the fitted values of all k studies simultaneously, we can calculate a measure analogous to Cook's distance [24], which is given in the meta-analytic context by

$$D_i = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})'(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}) \tag{20}$$

or equivalently

$$D_i = \sum \frac{(\hat{\mu}_i - \hat{\mu}_{i(-i)})^2}{v_i + \hat{\tau}^2}, \tag{21}$$

where $\hat{\boldsymbol{\beta}}_{(-i)}$ denotes the vector of parameter estimates from the fitted model after deletion of the ith study. Accordingly, a D_i value can be interpreted as the Mahalanobis distance between the entire set of predicted values once with the ith study included and once with the ith study excluded from the model fitting. Moreover, letting $\chi^2_{p',1-\alpha}$ denote the $100 \times (1-\alpha)$th percentile of a chi-square distribution with $p' = (p+1)$ degrees of freedom, note that the set of $\boldsymbol{\beta}$ values for which

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \chi^2_{p',1-\alpha} \tag{22}$$

defines a $100 \times (1-\alpha)$% joint confidence region for the p' regression coefficients in the model. Therefore, a value of D_i equal to $\chi^2_{p',1-\alpha}$ indicates that the deletion of the ith study would move the parameter estimates to the edge of a $100 \times (1-\alpha)$% joint confidence region based on the complete data. Following Cook and Weisberg [24], we may therefore consider values of D_i exceeding $\chi^2_{p',0.5}$ to warrant further inspection.

We can also directly examine the influence of deleting the ith case on each individual parameter estimate [26]. For this, we can calculate

$$\text{DFBETAS}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\sqrt{(\boldsymbol{X}'\tilde{\boldsymbol{W}}_{(-i)}\boldsymbol{X})^{-1}_{[j'j']}}}, \tag{23}$$

where $(\boldsymbol{X}'\tilde{\boldsymbol{W}}_{(-i)}\boldsymbol{X})^{-1}_{[j'j']}$ denotes the value of the $(j+1)$th diagonal element of the matrix $(\boldsymbol{X}'\tilde{\boldsymbol{W}}_{(-i)}\boldsymbol{X})^{-1}$ and $\tilde{\boldsymbol{W}}_{(-i)} = \text{diag}[1/(v_1+\hat{\tau}^2_{(-i)}), 1/(v_2+\hat{\tau}^2_{(-i)}), \ldots, 1/(v_k+\hat{\tau}^2_{(-i)})]$. In the linear regression context, values of DFBETAS_{ij} greater than 1 are often considered to indicate influential cases when analyzing small to medium data sets (e.g. [27]), a guideline that may also be useful for the meta-analytic context (where k generally tends to be small). Note that (23) simplifies to

$$\text{DFBETAS}_i = (\hat{\mu} - \hat{\mu}_{(-i)})\sqrt{\sum_{l=1}^{k} \tilde{w}_{l(-i)}} \tag{24}$$

in the random-effects model, where $\tilde{w}_{l(-i)} = 1/(v_l + \hat{\tau}^2_{(-i)})$. The DFBETAS_i statistic, therefore, formalizes the common practice of examining the change in the overall effect size estimate from a random-effects model when excluding each study in turn.

The influence of the ith study can also be examined by means of the change in the variance–covariance matrix of the parameter estimates when excluding the ith study from the model fitting [26]. For mixed-effects models, we can compute the ratio of the generalized variances, given by

$$\text{COVRATIO}_i = \frac{\det[\text{Var}[\hat{\boldsymbol{\beta}}_{(-i)}]]}{\det[\text{Var}[\hat{\boldsymbol{\beta}}]]}. \tag{25}$$

For random-effects models, this simplifies to

$$\text{COVRATIO}_i = \frac{\text{Var}[\hat{\mu}_{(-i)}]}{\text{Var}[\hat{\mu}]}. \tag{26}$$

A COVRATIO_i value below 1, therefore, indicates that removal of the ith study actually yields more precise estimates of the model coefficients (or equivalently, that addition of the ith study actually reduces precision).

Similarly, large changes in the estimate of τ^2 after exclusion of the ith study can signal the presence of outliers and/or influential cases. For example,

$$R_i = 100 \times (\hat{\tau}^2 - \hat{\tau}^2_{(-i)})/\hat{\tau}^2 \tag{27}$$

quantifies the change (in percent) in the estimate of τ^2 when the ith study is excluded relative to the estimated amount of (residual) heterogeneity when all the studies are included. Therefore, a large positive value of R_i indicates that the removal of the ith study leads to a large decrease in the amount of (residual) heterogeneity, which would be expected to occur if the ith study is indeed an outlier. Finally, Hedges and Olkin [2] suggested examining changes in the Q (or Q_E) statistic when excluding each study in turn. Generally speaking, a closer examination of a particular study is warranted if there are large relative changes in the various indices for that study when compared with the other studies.

5. Examples

The potential usefulness of the various diagnostic measures presented above will now be illustrated with three examples. The selected examples are different in various ways, for example, in terms of the effect size measure (i.e. the relative risk, standardized mean difference, and correlation coefficient), the number of effect sizes (i.e. 13, 26, and 61), and the meta-analytic model used (i.e. a random-effects model and two mixed-effects models, one with two and the other with three moderators). The examples demonstrate how the diagnostic measures can be applied in a variety of research settings.

5.1. BCG vaccine for tuberculosis

The first example concerns a set of 13 clinical trials examining the effectiveness of the bacillus Calmette–Guerin (BCG) vaccine for preventing tuberculosis [28]. For each of the 13 studies, Table I shows the number of tuberculosis positive (*TB*+) and negative (*TB*−) cases in the treated (i.e. vaccinated) and control (i.e. not vaccinated) groups. In addition, the publication year and the absolute latitude of the study location are indicated.

For each study, we can calculate the relative tuberculosis risk of the treated versus the control group with $RR_i = (a_i/n_i^T)/(c_i/n_i^C)$, where a_i and c_i are the number of TB+ cases in the treated and control groups, respectively, and n_i^T and n_i^C are the total number of subjects in the respective groups. For the meta-analysis, we use the log relative risk, $y_i = \log(RR_i)$, as the effect size measure, whose sampling variance is approximately equal to $v_i = (1/a_i) - (1/n_i^T) + (1/c_i) - (1/n_i^C)$ (e.g. [9]). Values of y_i below 0 indicate a lower tuberculosis risk for the vaccinated group.

We will consider a mixed-effects model for these data including publication year and latitude as potential moderators. Including publication year as a moderator provides information about potential changes in the effectiveness of the vaccine over time. The latitude of the study location may be considered a surrogate marker for the amount of non-pathogenic mycobacteria in the environment, which are more abundant closer to the equator and may provide a natural immunity to tuberculosis [28, 29].

Figure 1 shows each of the two moderators plotted against the log relative risk in the 13 studies. The point sizes are drawn proportional to $w_i = 1/v_i$ to emphasize differences in the precision of the estimates. Both figures suggest a trend, with higher

Table I. Results from 13 clinical trials examining the effectiveness of the bacillus Calmette–Guerin (BCG) vaccine for preventing tuberculosis.

				Treated		Control	
Trial	Author(s)	Year	Absolute latitude	TB+	TB−	TB+	TB−
1	Aronson	1948	44	4	119	11	128
2	Ferguson and Simes	1949	55	6	300	29	274
3	Rosenthal *et al.*	1960	42	3	228	11	209
4	Hart and Sutherland	1977	52	62	13 536	248	12 619
5	Frimodt-Moller *et al.*	1973	13	33	5036	47	5761
6	Stein and Aronson	1953	44	180	1361	372	1079
7	Vandiviere *et al.*	1973	19	8	2537	10	619
8	TPT Madras	1980	13	505	87 886	499	87 892
9	Coetzee and Berjak	1968	27	29	7470	45	7232
10	Rosenthal *et al.*	1961	42	17	1699	65	1600
11	Comstock *et al.*	1974	18	186	50 448	141	27 197
12	Comstock and Webster	1969	33	5	2493	3	2338
13	Comstock *et al.*	1976	33	27	16 886	29	17 825

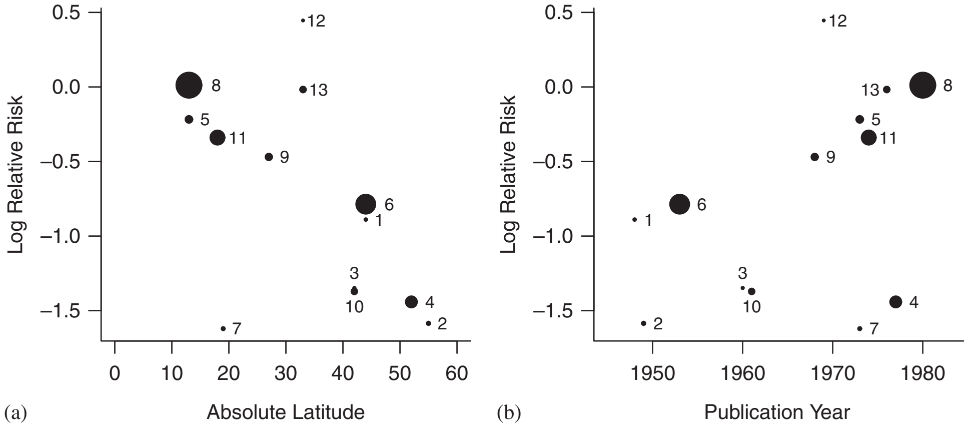

Figure 1. Plots of (a) absolute latitude and (b) publication year against the log relative risk for 13 clinical trials examining the effectiveness of the BCG vaccine for preventing tuberculosis (points drawn proportional to the inverse of the sampling variances).

vaccine effectiveness further away from the equator and decreasing effectiveness over time. However, some studies (i.e. studies 4, 7, 12, 13) appear to deviate from these trends.

To make the model intercept interpretable, latitude was centered at 33° and publication year at 1966 before fitting a mixed-effects model to these data. When fitting a mixed-effects model with both moderators included simultaneously, the estimated amount of residual heterogeneity is $\hat{\tau}^2 = 0.0790$ ($Q_E(\text{df} = 10) = 28.33, p < 0.002$). The estimated model coefficients (with 95% confidence intervals) are equal to $\hat{\beta}_0 = -0.71$ (-0.93 to -0.49) for the intercept (corresponding to the estimated average log relative risk at 33° absolute latitude in 1966), $\hat{\beta}_1 = -0.03$ (-0.05 to -0.01) for absolute latitude, and $\hat{\beta}_2 = 0.00$ (-0.03 to 0.03) for publication year. These results suggest that only absolute latitude is a significant moderator. Fitting two separate mixed-effects models for the two moderators leads to the same conclusion.

Figure 2 shows the hat values plotted against the studentized deleted residuals with point sizes drawn proportional to the Cook's distances. The figure illustrates that the Cook's distances essentially combine information about leverage and fit of a study. Most notable is the fourth study ($h_4 = 0.82$, $t_4 = -1.51$, $D_4 = 9.64$) with its very high leverage. However, while study 4 may be considered to be influential, it does not appear to be an outlier. On the other hand, studies 7 ($h_7 = 0.07$, $t_7 = -2.65$, $D_7 = 0.41$) and 13 ($h_{13} = 0.15, t_{13} = 2.06, D_{13} = 0.59$) have studentized deleted residuals larger than ±1.96, but are not nearly as influential. Here, it is also worth noting that removal of studies 7 and 13 leads to the largest reductions in the estimated amount of residual heterogeneity (i.e. by $R_7 = 43.24\%$ and $R_{13} = 26.18\%$, respectively), while removal of the influential fourth study yields a much smaller change in the estimate of τ^2 (i.e. by $R_4 = 14.50\%$).

The reason for the high leverage value of the fourth study becomes apparent when examining the moderator (i.e. $\boldsymbol{X}$) space directly. Figure 3 shows that the two moderators are in fact highly correlated ($r = -0.66$), with study 4 being a noticeable exception. As a result, its hat value will be large, making it a high leverage study.

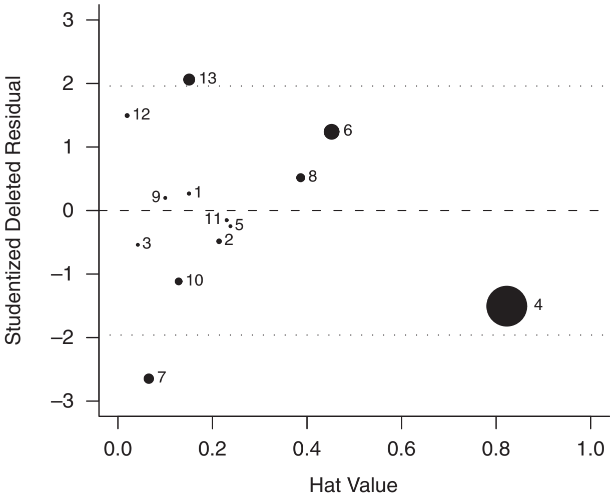

Figure 2. Plot of the hat values against the studentized deleted residuals for the BCG vaccine data (points drawn proportional to Cook's distances).

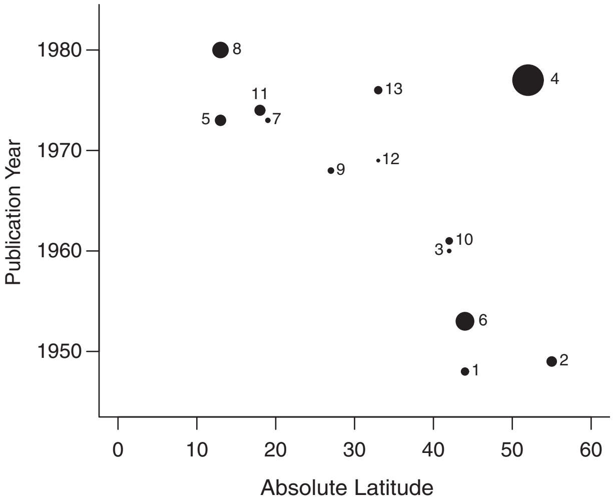

Figure 3. Plot of absolute latitude against publication year for the BCG vaccine data (points drawn proportional to the hat values).

The influence of the fourth study on the model coefficients is considerable, which is reflected in its large DFBETAS values ($\text{DFBETAS}_{4,0}=-0.87$, $\text{DFBETAS}_{4,1}=-2.86$, and $\text{DFBETAS}_{4,2}=-2.47$). In fact, neither of the two moderators reaches significance when the fourth study is removed from the dataset and the mixed-effects model is refitted. This is at least partly a result of the very high correlation between the two moderators, since each moderator examined individually in the context of a mixed-effects model is significant when the fourth study is removed. In summary then, these findings leave some doubt as to whether the varying prevalence of environmental mycobacteria at the study locations or whether changes over time (e.g. in respiratory hygiene) can account for the differences in the effectiveness of the BCG vaccine.

As a final note, the data can be used to illustrate how the externally studentized residual formalizes an outlier test under the mean shift outlier model. If we add a moderator to the mixed-effects model that is dummy coded equal to 1 for the fourth study and equal to 0 for the rest of the studies, then refitting the mixed-effects model yields $\hat{\beta}_3=-1.08$ with standard error equal to $\text{SE}[\hat{\beta}_3]=0.716$. The resulting test statistic is therefore $z=\hat{\beta}_3/SE[\hat{\beta}_3]=-1.51$, which corresponds exactly to the externally studentized residual for the fourth study given earlier (i.e. $t_4=-1.51$).

5.2. Writing-to-learn interventions

The second example is drawn from the educational research literature and concerns a meta-analysis regarding the effectiveness of writing-to-learn interventions (i.e. putting increased emphasis on writing assignments/exercises as part of the learning process) on academic achievement [30]. For illustration purposes, we will focus on a subset of the data, consisting of 26 studies conducted with high-school or college students (and leaving out one very old study from 1926).

Table II provides information about the publication year, whether the sample consisted of high-school or college students (dummy variable coded as college =1, high-school =0), the length of the intervention (in weeks), and whether the intervention

Table II. Results from 26 studies examining the effect of writing-to-learn interventions on academic achievement.

Study	Year	College	Length	Meta	y_i	v_i
1	1992	1	15	1	0.65	0.070
2	1994	1	9	0	−0.04	0.019
3	1996	1	1	0	0.03	0.009
4	1985	0	4	1	0.26	0.106
5	1986	1	4	0	0.06	0.040
6	1996	1	15	0	0.77	0.107
7	1994	1	15	1	0.00	0.021
8	1989	1	4	0	0.54	0.083
9	1996	1	14	0	0.20	0.086
10	1998	1	15	0	0.20	0.091
11	1991	1	4	0	−0.16	0.167
12	1985	1	3	0	0.51	0.065
13	1991	0	19	0	0.54	0.061
14	1993	0	12	1	0.37	0.060
15	1987	0	1	0	−0.13	0.037
16	1987	0	1	0	0.18	0.069
17	1993	0	1	0	0.27	0.018
18	1991	1	11	0	−0.32	0.060
19	1991	0	1	0	−0.12	0.023
20	1996	1	15	0	−0.07	0.033
21	1994	1	15	0	0.70	0.265
22	1987	1	2	1	0.49	0.039
23	1992	0	24	1	0.58	0.067
24	1980	1	15	0	0.63	0.168
25	1988	0	15	1	1.46	0.099
26	1989	1	15	0	0.25	0.072

incorporated prompts for ‘metacognitive reflection’ (dummy variable coded as yes $=1$, no $=0$) for each of the 26 studies. The effectiveness of the interventions was quantified in terms of the standardized mean difference (e.g. [8]),

$$d_i = \frac{\bar{x}_i^T - \bar{x}_i^C}{s_i}, \tag{28}$$

where $\bar{x}_i^T$ and $\bar{x}_i^C$ are the mean scores for the treatment and control groups on the academic achievement measure used in the ith study (e.g. grade, exam score), and s_i is the pooled standard deviation of the scores in the two groups. For the meta-analysis, we use $y_i = c(m_i)d_i$, where $m_i = n_i^T + n_i^C - 2$, n_i^T and n_i^C are the sample sizes of the two groups, and $c(m_i) = 1 - 3/(4m_i - 1)$ is a correction factor for the slight positive bias in the standardized mean difference [10]. The sampling variance of the standardized mean difference can be estimated with

$$v_i = \frac{n_i^T + n_i^C}{n_i^T n_i^C} + \frac{y_i^2}{2(n_i^T + n_i^C)}. \tag{29}$$

Since only the total sample size is reported in [30], we approximate v_i by assuming $n_i^T = n_i^C$, so $v_i \approx (8 + y_i^2)/(2N_i)$, where $N_i = n_i^T + n_i^C$ is the total sample size. Table II provides the y_i and v_i values for each study.

We will consider a mixed-effects model for these data with intervention length and the two dummy variables (metacognition and college). Figure 4 shows a plot of the effect size estimates as a function of intervention length. High-school and college samples are distinguished by the plotting symbol (circle versus square, respectively) with interventions prompting for metacognition using a filled symbol and non-prompting interventions a non-filled symbol. The figure shows that study 25 appears to have a much larger effect than the rest of the studies and may be an outlier.

The results from the mixed-effects model indicate a significant amount of residual heterogeneity ($\hat{\tau}^2 = 0.0472; Q_E(\text{df} = 22) = 44.14, p < 0.005$). The estimated model coefficients (with 95% confidence intervals) are equal to $\hat{\beta}_0 = 0.12$ (-0.13 to 0.37) for the intercept, $\hat{\beta}_1 = 0.01$ (-0.01 to 0.03) for intervention length, $\hat{\beta}_2 = 0.24$ (-0.06 to 0.54) for metacognition, and $\hat{\beta}_3 = -0.10$ (-0.37 to 0.16) for college. None of the estimated coefficients are, therefore, significantly different from 0 (i.e. all confidence intervals include the value 0).

Figure 5 shows the studentized deleted residuals, DFFITS values, Cook’s distances, and COVRATIO values for this model (note that the COVRATIO values are plotted on a log scale, so that deviations below and above 1 can be directly compared). Studies

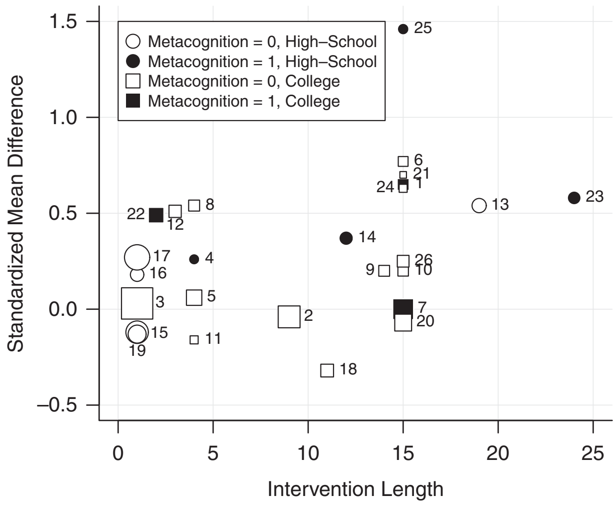

Figure 4. Estimated effects of writing-to-learn interventions on academic achievement as a function of intervention length in 26 studies.

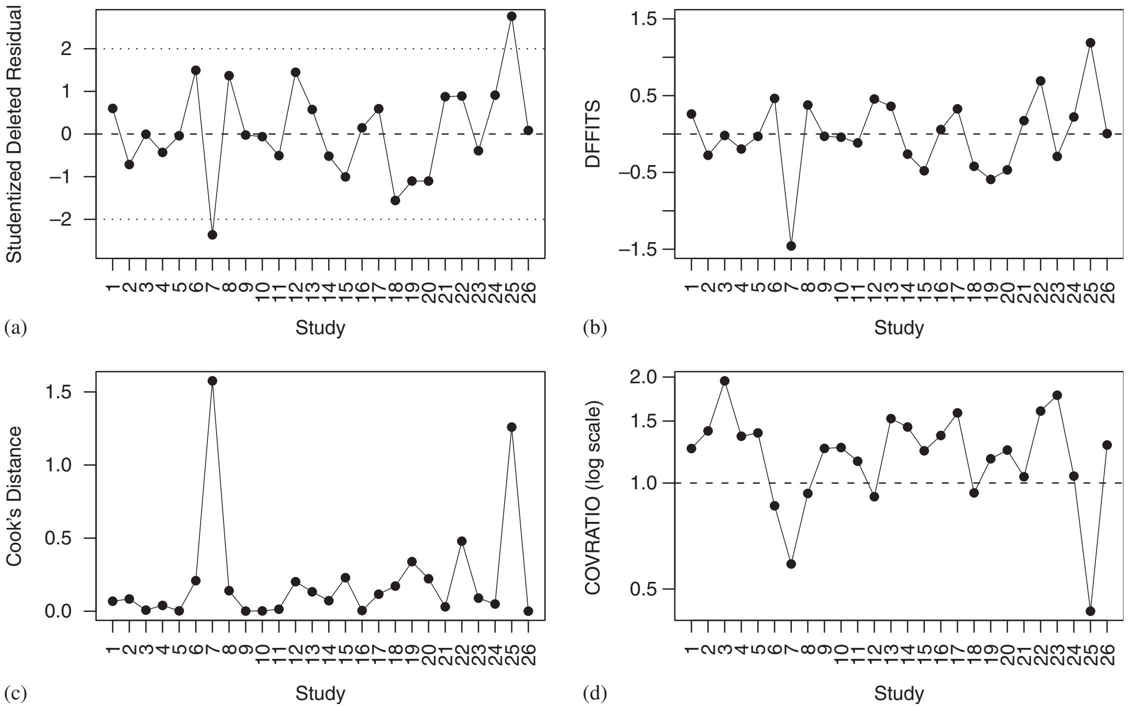

Figure 5. Plot of the (a) studentized deleted residuals; (b) DFFITS values; (c) Cook's distances; and (d) COVRATIO values for 26 studies examining the effectiveness of writing-to-learn interventions on academic achievement.

7 and 25 are identified as potential outliers and also appear to be influential cases. The COVRATIO values for these two studies also suggest that precision could be gained by their removal.

A reexamination of Figure 4 reveals that the standardized mean difference in study 7 is relatively low for an intervention prompting for metacognitive reflection. After removal of studies 7 and 25, the mixed-effects model suggests a significantly higher effect for interventions using metacognitive prompts ($\hat{\beta}_2=0.31$ with a 95% confidence interval from 0.03 to 0.59) and no significant amount of residual heterogeneity ($\hat{\tau}^2=0.0144$; $Q_E(\mathrm{df}=20)=26.07, p=0.16$). Since this conclusion about the influence of metacognitive prompts on the intervention effectiveness can only be reached after removing these two studies, it must be treated with caution.

5.3. Relationship between organizational commitment and job performance

The third example, coming from the business research literature, is based on a meta-analysis of the relationship between organizational commitment and salesperson job performance [31]. The meta-analysis comprises a total of 61 correlations based on 14 169 individuals. The example, therefore, demonstrates the use of the outlier and influence diagnostics when a large number of effects are included in a meta-analysis.

Before conducting the meta-analysis, the observed correlation coefficients, denoted by r_i, are transformed with Fisher's r-to-z transformation [11]. Therefore, we will use

$$y_i = \frac{1}{2}\ln\left[\frac{1+r_i}{1-r_i}\right] \tag{30}$$

as the effect size measure, with $v_i = 1/(n_i - 3)$ as the corresponding sampling variance. These values are provided in Table III.

The results from a random-effects model indicate a significant amount of between-study heterogeneity ($\hat{\tau}^2 = 0.0166$; $Q(\mathrm{df} = 60) = 285.65$, $p < 0.001$). The estimated value of $\hat{\mu}$ is equal to 0.19 with a 95% confidence interval from 0.15 to 0.23. Figure 6 shows the studentized deleted residuals, Cook's distances, and COVRATIO values for this model. Studies 8 and 56 are identified as both potential outliers and influential cases. The COVRATIO values for these two studies also suggest that precision could be gained by their removal. For example, without study 8, the estimate of $\hat{\mu}$ would be approximately 25% more efficient (i.e. the inverse of the COVRATIO value for the eighth study).

It is worth noting that none of the Cook's distances are actually larger than the 50th percentile of a chi-square distribution with 1 degree of freedom (i.e. $\chi^2_{1;0.5} = 0.45$). Nevertheless, it is clear from Figure 6 that two of the Cook's distances are comparatively large when compared with the other studies. Therefore, any of the aforementioned threshold values should be treated with some caution and should not replace an examination of the magnitudes of the various influence measures relative to each other.

Finally, this last example also demonstrates that the presence of potentially influential outliers may not necessarily call into question the conclusions from a meta-analysis. In particular, refitting the random-effects model after removal of studies 8 and 56 still leads to the finding that organizational commitment and job performance are positively (and significantly) correlated ($\hat{\mu} = 0.17$ with a 95% confidence interval from 0.14 to 0.20). Ensuring that the conclusions do not hinge on a few unusual studies therefore helps to demonstrate the robustness of the findings in this example.

6. Discussion

The main objective of this paper was to extend well-known outlier and influence diagnostics from standard linear regression to the meta-analytic context with particular emphasis on random- and mixed-effects models. The outlier and influence diagnostics presented in this paper are logical extensions of the corresponding measures in standard linear regression. In fact, when the sampling variances are homoscedastic (i.e. $v_i = v$ for all i) as assumed in standard linear regression, then all of the proposed measures simplify to the corresponding measures from standard linear regression[‡]. Moreover, the studentized deleted residual for random- and mixed-effects models given in this paper is the logical extension of this measure for fixed-effects models given by Hedges and Olkin [2].

We have used three examples to illustrate how these procedures could be applied to various research settings. The examples demonstrate that careful scrutiny of the effect sizes with the help of these procedures can yield important insights that either strengthen the conclusions from a meta-analysis or leave some doubts regarding their robustness.

While most researchers agree that it is necessary to examine the data for potential outliers and influential studies in a meta-analysis (e.g. [2–5]), Hunter and Schmidt [32] recommended against the use of outlier analyses in meta-analysis. The primary reason behind their position on outlier diagnostics is that 'it is almost impossible to distinguish between large sampling errors and true outliers (i.e. actual erroneous data)' (see [33], p. 110).

It is indeed true that unusually large or small effects could just be a result of chance alone. Therefore, the fact that an effect is particularly large or small should not by itself be taken as grounds for the routine deletion of the study reporting such an effect. However, it is important to emphasize that we are not advocating the routine deletion of outliers or influential studies. Instead, the statistics presented in the present paper should be used as part of sensitivity analyses. Some of the examples demonstrate the tentativeness of certain conclusions. When the removal of just one or two studies has a considerable impact on the conclusions from a meta-analysis, then these conclusions must be stated with some caution. If the outliers or influential cases do not alter the conclusions, researchers can be more confident that the meta-analytic findings are robust to outliers or influential cases.

As we are proposing various measures as tools to be used in the context of a sensitivity analysis, we have also omitted a discussion of methods to control the family-wise Type I error or false discovery rate (e.g. [34]) when using the internally and externally studentized residuals to screen for outliers. Unquestionably, a certain number of studies will have large residuals simply due to chance alone even when the model is correctly specified and using such methods would reduce the number of studies flagged as potential outliers. However, the goal is not to identify which studies are the 'real' outliers (essentially no method will be able to accomplish this whenever chance variation is assumed to be present in our estimates), but to examine whether we happen to be in the uncomfortable situation where the conclusions from the meta-analysis happen to depend on a few (potentially unusual) studies.

[‡] *The only exception is Cook's distance, which will differ by the factor p′. This difference arises as Cook's distance is derived in standard linear regression via its relationship to a joint confidence region defined by an F-distribution (e.g. [25, p. 119–120]), while the derivation used in the meta-analytic context is based on a joint confidence region defined by a chi-square-distribution (see Equation (22)).*

Table III. Results from 61 studies examining the relationship between organizational commitment and job performance.

Study	y_i	v_i
1	0.020	0.0057
2	0.121	0.0044
3	0.090	0.0018
4	0.203	0.0056
5	0.080	0.0065
6	0.040	0.0056
7	0.151	0.0040
8	0.485	0.0013
9	0.245	0.0047
10	0.234	0.0046
11	0.203	0.0088
12	0.234	0.0047
13	0.310	0.0046
14	0.332	0.0065
15	0.448	0.0128
16	0.040	0.0083
17	0.010	0.0233
18	0.151	0.0012
19	0.060	0.0020
20	0.141	0.0059
21	−0.010	0.0132
22	0.365	0.0069
23	0.030	0.0019
24	−0.050	0.0088
25	0.172	0.0044
26	0.266	0.0085
27	0.040	0.0049
28	0.255	0.0068
29	0.060	0.0019
30	0.090	0.0023
31	0.131	0.0033
32	0.310	0.0169
33	−0.161	0.0233
34	0.365	0.0130
35	0.131	0.0032
36	0.141	0.0015
37	0.192	0.0182
38	0.354	0.0036
39	−0.030	0.0161
40	0.310	0.0033
41	0.365	0.0076
42	0.203	0.0278
43	0.050	0.0047
44	0.172	0.0088
45	−0.060	0.0054
46	0.110	0.0108
47	0.576	0.0135
48	0.576	0.0137
49	0.354	0.0066
50	0.234	0.0078
51	0.203	0.0031
52	0.266	0.0019
53	0.080	0.0027
54	0.060	0.0094
55	0.050	0.0026
56	0.725	0.0122
57	0.536	0.0238
58	0.192	0.0115
59	0.090	0.0033
60	0.050	0.0189
61	0.348	0.0023

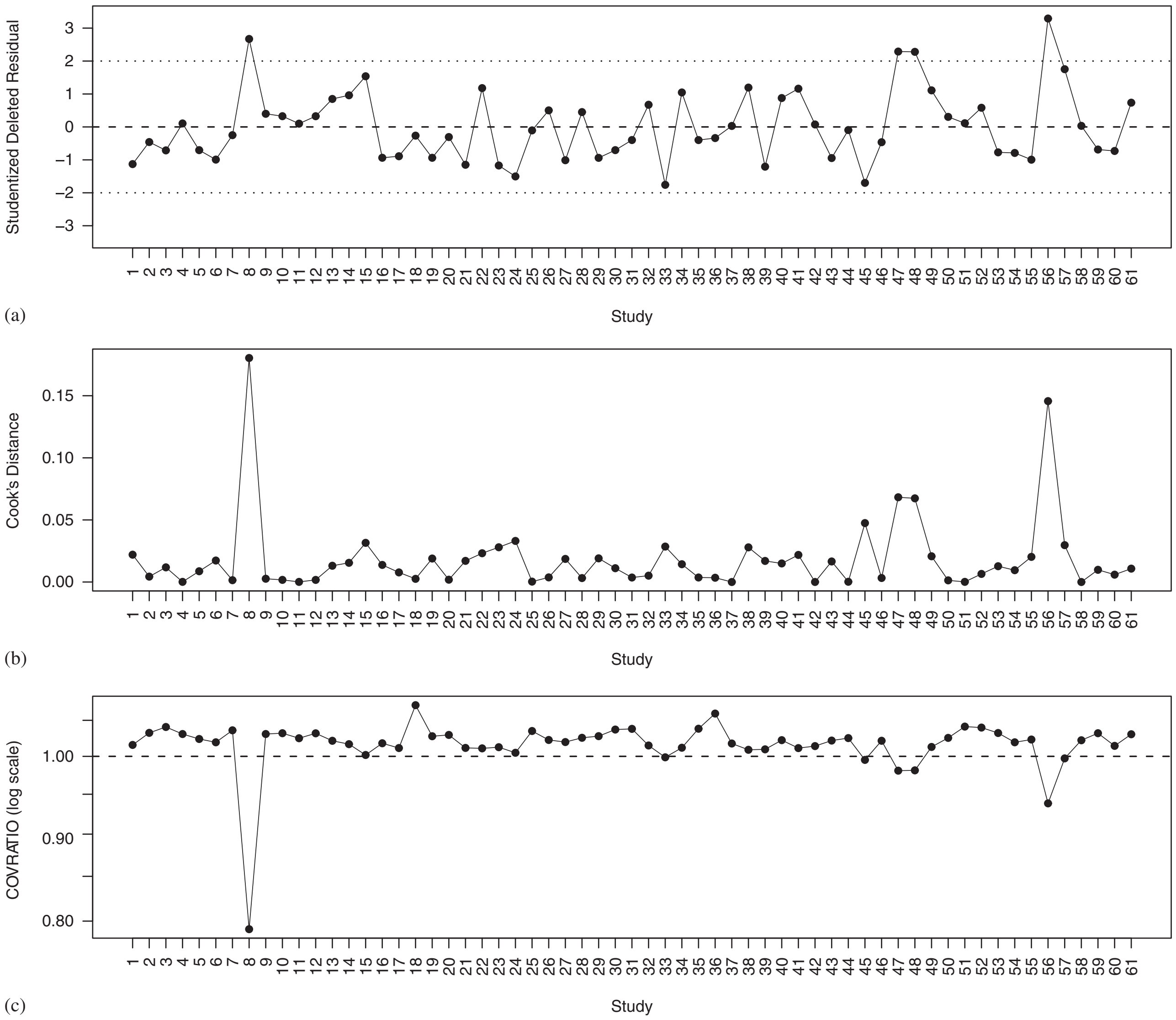

Figure 6. Plot of the (a) studentized deleted residuals; (b) Cook's distances; and (c) COVRATIO values for 61 studies examining the relationship between organizational commitment and job performance.

Moreover, studies identified as potential outliers should always be carefully scrutinized in terms of their contents. Outliers and influential cases can actually reveal patterns that may lead to new insights about study characteristics that could be acting as potential moderators [2, 3, 35]. For example, studies with unusually large or small effects can point to characteristics of treatments or experimental conditions that produce these types of effects [36]. In some cases, it may also be possible to corroborate the statistical identification of an outlier in other ways (e.g. when the effect size estimate was computed incorrectly or a moderator was miscoded). For example, a close examination of the influential fourth study from the BCG vaccine meta-analysis suggests that using the publication year as a moderator may not be the best choice for examining potential changes in the effectiveness of the vaccine over time. In particular, the fourth study was already started in 1950 with 89% of the tuberculosis cases actually occurring during the first 10 years of the study. Coding the year as 1977 for a relative risk that essentially reflects the data up to 1960 may have contributed to the fourth study becoming such an unusual case. Of course, such findings are exploratory and data driven and must also be treated with caution. Nevertheless, they may reveal potentially interesting areas for future research.

If the findings of a meta-analysis are sensitive to outliers or influential cases, another interesting direction to pursue during the analysis is to use models that are more robust in the presence of studies with large residuals. For example, models that allow for long/heavy-tailed and even skewed distributions for the random effects are proposed in [37–39]. The use of such distributions results in an automatic downweighting of outliers and therefore provides conclusions that are more robust in the presence of such cases.

Finally, it is worth noting that all the diagnostic measures presented in the present paper can be easily obtained with the *metafor* package (http://cran.r-project.org/package=metafor), an add-on package for conducting meta-analyses with the statistical software R (http://www.r-project.org). The package documentation and [40] describe how to fit the various meta-analytic models and how to obtain the diagnostic measures described in the current paper with the package. It is hoped that the availability of such software will support the routine use of the proposed methods in practice.

References

1. Cooper H, Hedges LV, Valentine JC. *The Handbook of Research Synthesis*, (2nd edn). The Russell Sage Foundation: New York, 2009.
2. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Academic Press: New York, 1985.
3. Light RJ, Pillemer DB. *Summing Up*: *The Science of Reviewing Research*. Harvard University Press: Cambridge, MA, 1984.
4. Lipsey MW, Wilson DB. *Practical Meta-Analysis*. Sage Publications: Thousand Oaks, CA, 2001.
5. Rosenthal R. Writing meta-analytic reviews. *Psychological Bulletin* 1995; **118**:183–192.
6. Greenhouse JB, Iyengar S. Sensitivity analysis and diagnostics. In *The Handbook of Research Synthesis and Meta-Analysis* (2nd edn), Cooper H, Hedges LV, Valentine JC (eds). The Russell Sage Foundation: New York, 2009; 417–433.
7. Wang MC, Bushman BJ. Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods* 1998; **3**:46–54.
8. Borenstein M. Effect sizes for continuous data. In *The Handbook of Research Synthesis and Meta-Analysis* (2nd edn), Cooper H, Hedges LV, Valentine JC (eds). The Russell Sage Foundation: New York, 2009; 221–235.
9. Fleiss JL, Berlin JA. Effect sizes for dichotomous data. In *The Handbook of Research Synthesis and Meta-Analysis* (2nd edn), Cooper H, Hedges LV, Valentine JC (eds). The Russell Sage Foundation: New York, 2009; 237–253.
10. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 1981; **6**:107–128.
11. Fisher RA. On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron* 1921; **1**:1–32.
12. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954; **10**:101–129.
13. Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis *Psychological Methods* 1998; **3**:486–504.
14. Hunter JE, Schmidt FL. Fixed effects vs. random effects meta-analysis models: implications for cumulative research knowledge. *International Journal of Selection and Assessment* 2000; **8**:275–292.
15. Shadish WR, Haddock CK. Combining estimates of effect size. In *The Handbook of Research Synthesis and Meta-Analysis* (2nd edn), Cooper H, Hedges LV, Valentine JC (eds). The Russell Sage Foundation: New York, 2009; 257–277.
16. Raudenbush SW. Analyzing effect sizes: random effects models. In *The Handbook of Research Synthesis and Meta-Analysis* (2nd edn), Cooper H, Hedges LV, Valentine JC (eds). The Russell Sage Foundation: New York, 2009; 295–315.
17. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random effects regression model for meta-analysis. *Statistics in Medicine* 1995; **14**:395–411.
18. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine* 2002; **21**:1559–1573.
19. Viechtbauer W. Bias efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* 2005; **30**:261–293.
20. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
21. Hartung J. An alternative method for meta-analysis. *Biometrical Journal* 1999; **41**:901–916.
22. Sánchez-Meca J, Marín-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods* 2008; **13**:31–48.
23. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* 2003; **22**:2693–2710.
24. Cook RD, Weisberg S. *Residuals and Influence in Regression*. Chapman and Hall: New York, 1982.
25. Weisberg S. *Applied Linear Regression* (2nd edn). Wiley: New York, 1985.
26. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics*. Wiley: New York, 1980.
27. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. *Applied Linear Statistical Models*. (4th edn). Irwin: Chicago, 1996.
28. Colditz GA, Brewer TF, Berkey CS, Wilson ME, Burdick E, Fineberg HV, Mosteller F. Efficacy of BCG vaccine in the prevention of tuberculosis: meta-analysis of the published literature. *Journal of the American Medical Association* 1994; **271**:698–702.
29. Ginsberg AM. The tuberculosis epidemic: scientific challenges and opportunities. *Public Health Reports* 1998; **113**:128–136.
30. Bangert-Drowns RL, Hurley MM, Wilkinson B. The effects of school-based writing-to-learn interventions on academic achievement: a meta-analysis. *Review of Educational Research* 2004; **74**:29–58.
31. Jaramilloa F, Mulkib JP, Marshall GW. A meta-analysis of the relationship between organizational commitment and salesperson job performance: 25 years of research. *Journal of Business Research* 2005; **58**:705–714.
32. Hunter JE, Schmidt FL. *Methods of Meta-Analysis*: *Correcting Error and Bias in Research Findings* (2nd edn). Sage: Thousand Oaks, CA, 2004.
33. Schmidt FL. Meta-analysis: a constantly evolving research integration tool. *Organizational Research Methods* 2008; **11**:96–113.
34. Shaffer JP. Multiple hypothesis testing. *Annual Review of Psychology* 1995; **46**:561–584.
35. Hedges LV. Issues in meta-analysis. *Review of Research in Education* 1986; **13**:353–398.
36. Mosteller F, Colditz GA. Understanding research synthesis (meta-analysis). *Annual Reviews of Public Health* 1996; **17**:1–23.
37. Demidenko E. *Mixed Models*: *Theory and Applications*. Wiley: Hoboken, NJ, 2004.
38. Lee KJ, Thompson SG. Flexible parametric models for random-effects distributions. *Statistics in Medicine* 2008; **27**:418–434.
39. Baker R, Jackson D. A new approach to outliers in meta-analysis. *Health Care Management Science* 2008; **11**:121–131.
40. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 2010; **36**:1–48.