Accounting for Heterogeneity via Random-Effects Models and Moderator Analyses in Meta-Analysis

Wolfgang Viechtbauer

University of Maastricht, The Netherlands

Abstract. To conduct a meta-analysis, one needs to express the results from a set of related studies in terms of an outcome measure, such as a standardized mean difference, correlation coefficient, or odds ratio. The observed outcome from a single study will differ from the true value of the outcome measure because of sampling variability. The observed outcomes from a set of related studies measuring the same outcome will, therefore, not coincide. However, one often finds that the observed outcomes differ more from each other than would be expected based on sampling variability alone. A likely explanation for this phenomenon is that the true values of the outcome measure are heterogeneous. One way to account for the heterogeneity is to assume that the heterogeneity is entirely random. Another approach is to examine whether the heterogeneity in the outcomes can be accounted for, at least in part, by a set of study-level variables describing the methods, procedures, and samples used in the different studies. The purpose of the present paper is to discuss these different approaches with particular emphasis on the interpretation of the results and practical issues.

Keywords: meta-analysis, moderator analysis, random-effects model, meta-regression

Introduction

It has been estimated that the number of journals doubles every 15 years (Meadows, 1998) and given the corresponding increases in journal sizes and publication frequencies, the growth in the number of scientific papers published every year can be assumed to take on even more staggering proportions. While electronic databases and full-text electronic versions of journals have made it easier to access and maintain awareness of the relevant literature, reading and processing the existing literature is becoming increasingly difficult.

To address this problem, statistical methods have been developed over the past three to four decades to facilitate the systematic review of the literature (Chalmers, Hedges, & Cooper, 2002). A systematic literature review conducted with the aid of such methods is called a meta-analysis, a term that was coined by Gene Glass in 1976. The availability of meta-analyses may greatly reduce the amount of time and effort required by researchers and practitioners to stay updated on the research within their field, since reading and processing a single well-conducted and thorough literature review on a particular topic takes less time and effort than reading a dozen or sometimes hundreds of articles on the same issue. Moreover, meta-analyses have been argued to provide higher quality information in comparison with narrative literature reviews, which may be more subjective, inefficient, and selective in their scope (e.g., Jackson, 1980; Light & Pillemer, 1984). Given the increasing rate at which meta-analyses are being published in various fields since the beginning of the 1980s (Lee, Bausell, & Berman, 2001; Schulze, 2004), it appears that researchers have been eager to assimilate this new technique into their methodological repertoire.

To conduct a meta-analysis, the relevant outcome of each study to be included needs to be summarized in such a way that the results from the different studies are expressed on a common scale (e.g., Fleiss, 1994; Rosenthal, 1994). Typically, the outcome of interest is some measure of effect, association, or the size of a group difference (e.g., the effectiveness of a treatment, the degree of association between two constructs, or the amount by which two groups differ with respect to some characteristic or attribute). These values then become the data for further analysis, such as in the estimation of an overall effect, association, or group difference (e.g., Hedges, 1982). However, empirical evidence suggests that the effect of a treatment, the strength of an association, or the size of a group difference is often not a single unchanging value, but may actually differ across studies (Field, 2005; Higgins, Thompson, Deeks, & Altman, 2003; Lipsey & Wilson, 2001b). This raises the question of how to account for the variability in the outcomes across studies.

One option is to assume that such differences are entirely random (Hedges, 1983). An alternative approach is to examine whether at least part of the variability in the outcomes can actually be accounted for by systematic differences between the characteristics of the studies from which the outcomes have been derived (e.g., Pillemer & Light,

Zeitschrift für Psychologie / Journal of Psychology 2007; Vol. 215(2):104–121
DOI 10.1027/0044-3409.215.2.104

Table 1. Results from 17 studies comparing the effectiveness of St. John's wort with placebos for treating depression (Linde et al., 2005)

Study	m_i^T	n_i^T	m_i^C	n_i^C	RR_i	y_i	v_i	Weekly dosage	Only major depression	Baseline score	Duration in weeks
1	20	25	11	25	1.82	0.60	.061	2.66	No	19.5	8
2	14	20	9	20	1.56	0.44	.083	6.30	No	12.5	4
3	4	25	2	25	2.00	0.69	.670	6.30	Yes	22.7	4
4	20	32	6	33	3.44	1.23	.155	6.30	No	16.5	6
5	28	50	13	55	2.37	0.86	.074	6.30	No	15.8	4
6	34	48	25	49	1.39	0.33	.028	1.68	Yes	23.6	6
7	35	53	12	54	2.97	1.09	.075	6.30	Yes	20.7	4
8	24	49	16	49	1.50	0.41	.063	6.30	Yes	21.1	6
9	45	80	12	79	3.70	1.31	.080	3.50	Yes	19.4	6
10	67	106	22	47	1.35	0.30	.030	7.35	Yes	22.7	6
11	34	60	17	59	1.97	0.68	.055	6.30	No	16.7	6
12	46	70	34	70	1.35	0.30	.023	3.50	Yes	20.9	6
13	55	123	57	124	0.97	–0.03	.020	6.30	Yes	21.5	6
14	23	37	15	35	1.45	0.37	.055	6.30	Yes	19.9	6
15	26	98	19	102	1.42	0.35	.071	7.35[1]	Yes	22.5	8
16	46	113	56	116	0.84	–0.17	.022	8.40[1]	Yes	22.9	8
17	98	186	80	189	1.24	0.22	.012	6.30	Yes	21.9	6

Notes: n_i^T and n_i^C = number of participants in the treatment and the placebo group; m_i^T and m_i^C = number of participants with significant improvements between baseline and the follow-up assessment in the treatment and the placebo group; RR_i = relative improvement rate; y_i = log of the relative rate; v_i = estimated sampling variance of the log relative rate (see text for equations). [1] The value given is the midpoint of a range of different dosages used in the study.

1980). The purpose of the present paper is to discuss these different explanations with particular emphasis on the second approach, not only because it can lead to more rich and interesting results, but also because it is likely to be a more accurate reflection of the true state of affairs in many situations (Lipsey & Wilson, 2001b).

Meta-Analysis Example

A recently published meta-analysis will be used as an example throughout this article to provide a concrete backdrop for the discussion and to illustrate the methods and issues discussed. The example concerns the effectiveness of St. John's wort for treating depression. St. John's wort, extracted from the yellow-flowering herb *Hypericum perforatum*, has long been regarded as a safe and effective treatment for depression, but clinical studies comparing St. John's wort with placebo or standard antidepressant treatment have yielded mixed findings. Linde, Berner, Egger, and Mulrow (2005) recently conducted a meta-analysis of double-blind randomized controlled studies to provide a clarification of the evidence regarding the effectiveness of St. John's wort.

Table 1 shows the results from $k = 17$ studies, comparing St. John's wort with placebos. Given are the number of participants in the treatment and the placebo group (n_i^T and n_i^C, respectively), the number of participants who showed significant improvements in their condition between baseline and the follow-up assessment in the two experimental conditions (m_i^T and m_i^C, respectively), the relative rate of improvement (the improvement rate in the treatment group divided by the improvement rate in the placebo group), i.e.,

$$RR_i = (m_i^T/n_i^T)(m_i^C/n_i^C),$$

the log of the relative rate, i.e.,

$$y_i = \ln(RR_i),$$

and the estimated amount of sampling variability in the log relative rate, i.e.,

$$v_i = 1/m_i^T - 1/n_i^T + 1/m_i^C - 1/n_i^C.$$

A relative rate of 1 (or a log relative rate of 0) indicates equal improvement rates for the St. John's wort and the placebo group (i.e., St. John's wort is no more effective than placebos), a relative rate greater than 1 (or a log relative rate greater than 0) indicates a higher rate of improvement for the group receiving St. John's wort (i.e., St. John's wort is more effective than placebos), and a relative rate below 1 (or a log relative rate below 0) indicates a higher rate of improvement in the placebo group (i.e., St. John's wort is less effective than placebos).

Letting π_i^T and π_i^C denote the true but unknown probabilities of an improvement in the treatment and placebo

group in the i^{th} study, then $\theta_i = \ln(\pi_i^T/\pi_i^C)$ denotes the corresponding true but unknown log relative rate. The observed log relative rate y_i is a consistent and approximately normally distributed estimator of the true log relative rate θ_i with estimated variance equal to v_i (Fleiss, 1994)[1].

Also listed in Table 1 are the weekly dosage (in grams) of the *Hypericum* extract used in each study, whether a study was restricted to participants with major depression or not, the average score on the Hamilton Rating Scale for Depression (HRSD) at baseline (i.e., before treatment begin), and the number of treatment weeks before response assessment. The relevance of these variables will be discussed later on. Because of space considerations and for didactic purposes, only a subset of the complete dataset from Linde et al. (2005) is presented here. No substantive interpretation should, therefore, be attached to the results of the analyses given later.

It should be emphasized that the statistical methods discussed in this article are not restricted to meta-analyses using the (log) relative rate as the outcome measure of choice. Other outcome measures frequently used in meta-analyses are the standardized mean difference (the mean difference between two groups divided by the pooled standard deviation), the correlation coefficient (either in its raw form or after applying Fisher's variance stabilizing *z*-transformation), and the odds ratio (for more details on these and other outcome measures used in meta-analyses, see Fleiss, 1994, and Rosenthal, 1994).

Therefore, regardless of the outcome measure of choice, let y_i denote the observed outcome and θ_i the corresponding true but unknown value of the outcome measure for the i^{th} study. Depending on the outcome measure and context, it may be appropriate to call y_i a measure of "effect" or a measure of "association." For example, the relative rate could be regarded as an effect measure in the St. John wort meta-analysis (i.e., y_i indicates the effect of the treatment on the probability of improvement relative to that of a placebo group). The standardized mean difference is also usually interpreted as an effect measure, such as in meta-analyses of studies that compare two experimental or naturally defined groups with respect to some attribute assessed on a continuous scale (e.g., differences between men and women with respect to their risk-taking tendencies; see Byrnes, Miller, & Schafer, 1999). On the other hand, the correlation coefficient is typically used as a measure of (linear) association between two variables (e.g., the validity of employment interviews as predictors of actual job performance; see McDaniel, Whetzel, Schmidt, & Maurer, 1994).

The Homogeneous Situation

When presented with the results from a collection of studies to be included in a meta-analysis, the first thing we may notice is the fact that the outcomes from the various studies seldom provide a unanimous picture with respect to the strength of the effect or association. For example, in the St. John's wort meta-analysis (Table 1), the observed relative rates range from 0.84 to 3.70 with two studies yielding a relative rate below 1, 11 studies yielding a relative rate between 1 and 2, and four studies yielding a relative rate higher than 2. Therefore, a few studies suggest that St. John's wort may actually reduce improvement chances, a good number of studies indicate small to medium-sized benefits when taking St. John's wort, and a handful of studies provide evidence of more substantial benefits beyond those of a placebo effect.

However, even studies conducted under identical or nearly identical conditions may yield different and sometimes contradictory conclusions. Consider, for example, Study 13 (by Montgomery, Hübner, & Grigoleit, 2000) and Study 17 (by Lecrubier, Clerc, Didi, & Kieser, 1994) in Table 1. Both studies used the same weekly dosage of St. John's wort (900 mg), both assessed the treatment response after 6 weeks, both were restricted to participants suffering from major depression, the average baseline HRSD scores of the participants were almost identical in the two studies (21.5 and 21.9, respectively), and yet the first study found a relative rate of 0.97, while the other found a relative rate of 1.24. In fact, a 95% confidence interval for the true relative rate yields the interval bounds (0.74, 1.28) for the first and (1.01, 1.54) for the second study, once leading to the conclusion that St. John's wort is not significantly better than placebos (the value 1 is included in the confidence interval) and once leading to the conclusion that St. John's wort does provide benefits beyond those offered by placebos (the value 1 is not included in the interval)[2]. While these two studies actually yield conflicting conclusions, it may be the case that the true (log) relative rates are exactly the same in both studies and that the difference in the observed relative rates (and the corresponding difference in the statistical significance of the findings) is simply a result of random sampling fluctuations.

In general, the true effectiveness or association may be exactly the same ("homogeneous") for all k studies included in a meta-analysis (i.e., $\theta_i = \theta$ for $i = 1, \ldots, k$) so that differences between the observed outcomes (i.e., differences between the y_i values) would be a result of sampling fluctuations alone (Hedges & Vevea, 1998). While each

[1] The methods discussed throughout this paper are based on the assumption that the outcome measure used in the meta-analysis is (at least approximately) normally distributed. Taking the log of the relative rates greatly helps to improve the approximation to the normal distribution for this outcome measure.

[2] Since it is approximately true that $y_i|\theta_i \sim N(\theta_i, v_i)$, where y_i denotes the observed log relative risk, v_i the estimated sampling variance of y_i, and θ_i the true log relative risk for the i^{th} study, it follows that $y_i \pm 1.96\sqrt{v_i}$ gives the bounds of an approximate 95% confidence interval for the true log relative risk. Exponentiating those bounds yields an approximate 95% confidence interval for the true relative risk in the i^{th} study.

observed outcome then provides an estimate of the common θ, a more precise estimate of θ (e.g., the true effectiveness of a treatment or the true association between two variables) can be easily obtained simply by averaging the results from the studies. For example, the average of the log relative rates from Table 1 is equal to 0.53, which corresponds to a relative rate of exp(0.53) = 1.70, suggesting that the improvement rate is 1.7 times higher when treated with St. John's wort than when receiving placebos.

A more refined analysis takes into consideration the fact that the studies included in a meta-analysis are typically not of equal size. For example, the smallest study (Study 2) yielded a relative rate of 1.56 based on a total of 40 participants, while the largest study (Study 17) yielded a relative rate of 1.24 based on a total of 375 participants. The difference in sample size could be incorporated into the analysis by giving more weight to the latter result, since studies with larger sample sizes tend to provide more accurate estimates of θ. In other words, all else being equal, the value of y_i will tend to be closer to θ for a larger study than for a smaller study, which is reflected by the fact that the amount of sampling variability in y_i tends to decrease as the total sample size increases (e.g., the estimated amount of sampling variability is almost seven times larger for Study 2 (v_2 = .083) than for Study 17 (v_{17} = .012)). Since the amount of sampling variability indicates the degree of imprecision in an estimate, the common practice in meta-analysis is to estimate θ by calculating a weighted average of the observed outcomes with

$$\hat{\theta} = (\Sigma w_i y_i)/\Sigma w_i, \quad (1)$$

using weights equal to the inverse of the estimated sampling variances (i.e., $w_i = 1/v_i$) as indicators of the precision in the estimates (Hedges, 1982). An approximate 95% confidence interval for θ can then be obtained with

$$\hat{\theta} \pm 1.96\sqrt{1/\Sigma w_i}\ . \quad (2)$$

For the St. John's wort data, this approach yields an estimate of $\hat{\theta}$ = 0.33 for the log relative rate, corresponding to a relative rate of exp(0.33) = 1.39. The bounds of the 95% confidence interval for the log relative rate are given by (0.23, 0.42). Converted back to relative rates, the confidence interval bounds are equal to (1.26, 1.52). Note that the confidence interval does not include the value 1, suggesting that the improvement rate is significantly higher with St. John's wort treatment than when receiving placebos.

Testing for the Presence of Heterogeneity

The assumption that the $\hat{\theta}_i$ values are homogeneous across studies (so that differences among the observed outcomes are a result of sampling variability alone) can actually be tested by calculating

$$Q = \Sigma w_i (y_i - \hat{\theta})^2, \quad (3)$$

which we compare against the upper one-sided critical value of a χ^2 distribution with k –1 degrees of freedom (e.g., Hedges, 1982, 1983). When Q exceeds the critical value, then this suggests that the observed outcomes differ more from each other than would be expected based on sampling variability alone. This is, in fact, what we would expect to observe if the θ_i values are not all equal to each other (i.e., if the θ_i values are "heterogeneous").

For the St. John's wort meta-analysis, we find that Q = 51.54, which is clearly larger than 26.30, the one-sided critical χ^2 value for α = .05 and 16 degrees of freedom. Therefore, we conclude that the true (log) relative rates are not the same for all of the studies (i.e., the true (log) relative rates are heterogeneous).

The Heterogeneous Situation

One way to explain the presence of heterogeneity is to assume that differences among the observed outcomes are not only a result of random sampling fluctuations, but are also caused by random variability at the study level (Hedges, 1983; Hedges & Vevea, 1998). To clarify this idea, it is helpful to think of the mechanism leading to the observed outcome for a particular study as a two-stage hierarchical process (see Figure 1).

At the first stage, the true effectiveness of a treatment or the true association between two variables for a particular study (i.e., a study's θ_i value) is assumed to be determined by a more or less countless number of unknown factors specifying, for example, how a study was designed, how it was carried out, and the circumstances under which it was conducted. Each factor taken by itself may only have a small influence on the true effectiveness or association strength, but in sum these factors yield a distribution of potential θ_i value. The actual characteristics of a particular study then give rise to a random draw from that distribution of true effects or association strengths, yielding a specific θ_i value. The form of the distribution is typically assumed to be normal, partly for convenience sake (computations are greatly simplified under this assumption), but also based on a consideration of the central limit theorem (i.e., the sum of a large number of independent variables tends toward a normal distribution under certain conditions).

Given the specific value of θ_i for a particular study, the second stage then concerns the random variability in the outcome y_i that arises out of the sampling of subjects. In other words, given θ_i for a particular study, there exists an entire distribution of potential y_i values that we could observe (i.e., the sampling distribution of y_i), but the selection of an actual sample then yields a specific observed out-

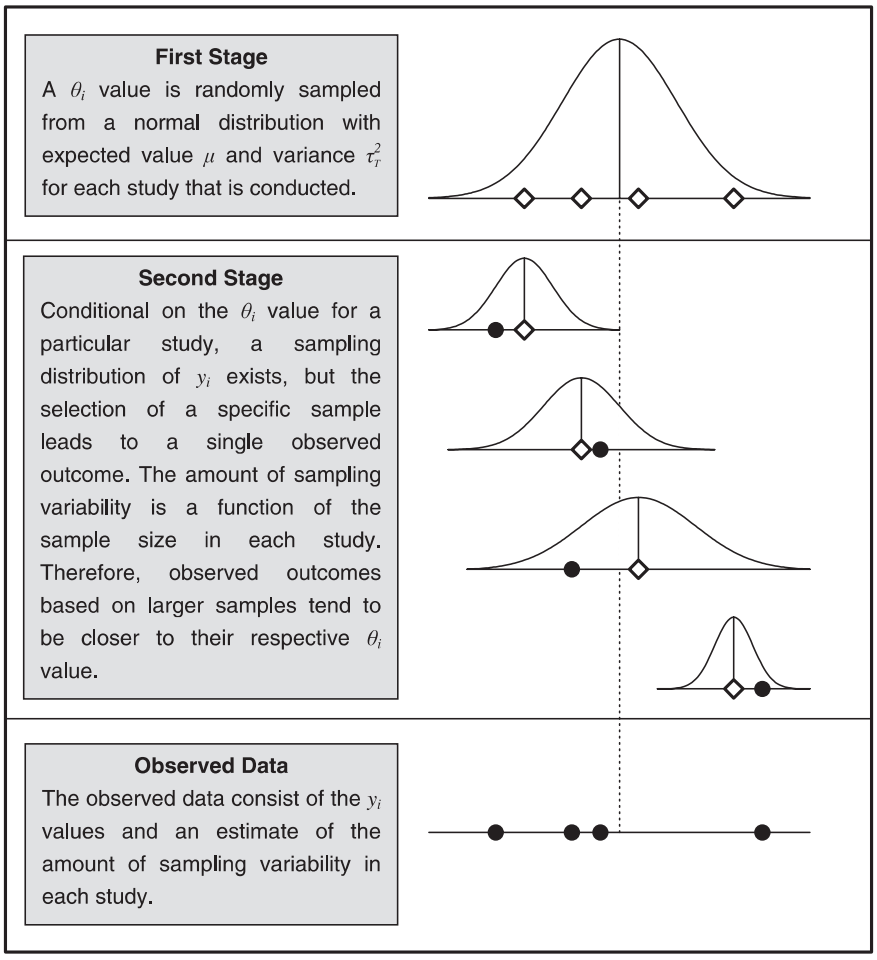

Figure 1. Schematic diagram of the random-effects model illustrating the two-step process leading to the observed outcomes.

come. For most outcome measures used in meta-analyses, the sampling distribution of y_i is assumed to be normal, although this is usually just an approximation that gets more accurate as the sample sizes within the studies become large. Moreover, larger sample sizes will yield y_i values that are, on average, closer to their respective θ_i values than smaller sample sizes.

In summary, we note that the observed outcomes are influenced by two sources of variability under this model: heterogeneity among the θ_i values at the first stage and sampling variability within the y_i values at the second stage, where only the latter is a function of the sample size within the studies. In other words, if the sample size within each study would be very large, then sampling variability essentially becomes negligible in all of the studies (i.e., $v_i \approx 0$) and, therefore, $y_i \approx \theta_i$. Nevertheless, we still would not expect all of the observed y_i values to coincide, since the corresponding θ_i values differ from each other due to random heterogeneity.

Under this so-called random-effects model, it is no longer appropriate to speak of "the" effect of a treatment or "the" association between two variables, since θ_i is no longer constant across studies. In fact, θ_i may be positive in some studies (e.g., corresponding to an effective treatment or a positive correlation) and negative in others (e.g., corresponding to a harmful treatment or a negative correlation). However, if θ_i really differs randomly across studies, then an appropriate question would be to ask about the average effect or association within the distribution of θ_i values. For example, for any particular study, St. John's wort may be an effective treatment ($\theta_i > 0$), no better than placebos ($\theta_i = 0$), or even a harmful treatment ($\theta_i < 0$), but instead of examining whether St. John's wort is an effective treatment in all situations, we can now try to determine whether St. John's wort is an effective treatment on average.

The analysis proceeds as follows. First, we estimate the amount of variability among the θ_i values, which will be denoted by τ^2_T. In other words, τ^2_T denotes the amount of heterogeneity within the distribution of true effects or associations, or more precisely, the variance of the random variable producing the θ_i values. An estimate of τ^2_T can be obtained with the estimator suggested by DerSimonian and Laird (1986), which is given by

$$\hat{\tau}^2_T = \frac{Q-(k-1)}{c}, \tag{4}$$

with Q as given earlier and

$$c = \Sigma w_i - \frac{\Sigma w_i^2}{\Sigma w_i}$$

(a negative value of $\hat{\tau}_T^2$ is truncated to zero). Next, we calculate new weights that are equal to $w_i^* = 1/(v_i + \hat{\tau}_T^2)$ and then estimate μ, the average (or better: the expected) value of θ_i in the distribution of θ_i values with

$$\hat{\mu} = (\Sigma w_i^* y_i)/\Sigma w_i^*. \quad (5)$$

Finally, an approximate 95% confidence interval for μ can be obtained with

$$\hat{\mu} \pm 1.96\sqrt{1/\Sigma w_i^*}. \quad (6)$$

Note that when τ_T^2 is estimated to be zero, then this suggests the absence of heterogeneity, in which case $\hat{\mu}$ and $\hat{\theta}$ as well as the respective confidence intervals are identical.

For the St. John's wort meta-analysis, we find $\hat{\tau}_T^2 = .091$ and $\hat{\mu} = 0.45$. Therefore, the average relative rate is estimated to be equal to exp(0.45) = 1.57, indicating that the improvement rate is, on average, almost 1.6 times higher for those receiving St. John's wort than those receiving placebos. The bounds of the 95% confidence interval for μ are equal to (0.27, 0.64), corresponding to relative rates of (1.31, 1.90). Given that the value 1 does not fall inside the confidence interval, we conclude that the average relative rate is significantly higher than 1, indicating that treatment with St. John's wort is, on average, effective[3].

However, it needs to be emphasized again that, under the assumptions of the random-effects model, the true log relative rate for any particular study is assumed to differ randomly from μ and may indeed be negative in some cases (in which case St. John's wort would actually reduce the chances of an improvement compared to placebos). If, indeed, the θ_i values are normally distributed and the estimates $\hat{\mu} = 0.45$ and $\hat{\tau}_T^2 = .091$ are assumed to be free of error, then we would expect 95% of the true log relative rates in studies examining the effectiveness of St. John's wort to fall between $0.45 - 1.96\sqrt{.091} = -0.14$ and $0.45 + 1.96\sqrt{.091} = 1.04$, which corresponds to the values (0.87, 2.83) in terms of relative rates[4]. Therefore, in some studies, the relative rate would be expected to fall below 1. In fact, from the given information it is easy to calculate that the true relative rate would be expected to fall below 1 in roughly 7% of all studies.

The Influence of Moderators

Empirical evidence suggests that heterogeneity is present in approximately 50% to 75% of all meta-analyses (Field, 2005; Higgins et al., 2003; Lipsey & Wilson, 2001b). A random-effects model analysis, which assumes that the heterogeneity among the θ_i values is completely unsystematic, is one approach to account for the presence of heterogeneity. However, it is possible, and in some cases rather likely, that the variability among the θ_i values is not entirely random, but actually quite systematic.

For example, the intensity of a treatment or intervention (expressed by study-level characteristics such as medication dosage or intervention length) may influence the outcome in a rather predictable manner. Suppose, for instance, that the treatment effectiveness is (approximately) linearly related to the intervention length. Therefore, the true outcome in the i^{th} study is given by $\theta_i = \beta_0 + \beta_1 x_i$, where x_i denotes the intervention length (in weeks) for the i^{th} study, β_0 denotes the treatment effectiveness when $x_i = 0$, and β_1 denotes how much the treatment effectiveness increases for a 1 week increase in intervention length. If the studies included in the meta-analysis differ with respect to the intervention length used, then the heterogeneity among the θ_i values would not be random, but actually systematic.

The undifferentiated application of a random-effects model to a situation where the heterogeneity among the θ_i values is systematic can actually lead to misleading or even nonsensical results. To illustrate this point, consider the following example. In certain contexts, evidence suggests that θ_i differs systematically for published journal articles and dissertations (e.g., Smith, 1980)[5]. Suppose now that this is indeed the case for a particular meta-analysis aggregating the results from a series of journal articles and dissertations that report the correlation between two constructs of interest. Then the true correlation in the i^{th} study is given by $\theta_i = \beta_0 + \beta_1 x_i$, with $x_i = 0$ for correlations reported in dissertations and $x_i = 1$ for correlations reported in journal articles. Consequently, β_0 denotes the true value of the correlation for dissertations and $\beta_0 + \beta_1$ denotes the true value of the correlation for journal articles. However, applying a random-effects model to such data, which assumes that the heterogeneity in the correlations is entirely random, would yield an estimate of μ that

[3] An inference about μ is an unconditional inference about the expected value of θ_i in the entire distribution of θ_i values. An alternative approach is to make an inference about the average effect or association in the specific set of studies included in the meta-analysis (i.e., $\bar{\theta} = \Sigma\theta/k$). The so-called conditional inference model applies in this case (e.g., Hedges & Vevea, 1998) and all of the results given for the homogeneous situation are correct, except that $\hat{\theta}$ is then an estimate of $\bar{\theta}$.

see note at the bottom

[4] The interval $\hat{\mu} \pm 1.96\sqrt{\hat{\tau}_T^2}$ is a so-called 95% credibility interval, as described by Hunter and Schmidt (1990). Assuming that the estimates of μ and τ_T^2 are free of error and that the θ_i values are normally distributed, the 95% credibility interval indicates the range of values where 95% of the θ_i values are expected to fall.

[5] Smith (1980) introduced the term "publication bias" to describe the phenomenon that the more accessible literature (e.g., published journal articles) may differ in systematic ways from less accessible works (e.g., theses, dissertations, unpublished articles), which, in turn, may bias the results from a meta-analysis that only focuses on the published literature. A more thorough discussion of this issue is beyond the scope of the present article, but the interested reader can consult, for example, Rothstein, Sutton, and Borenstein (2005).

This is not 100% accurate. If we use equation (1) under heterogeneity, then we are making a conditional inference about the **weighted** average of the θi values (i.e., about Σwiθi/Σw). If we want to make an inference about the unweighted average (i.e., about Σθi/k), then we must use an unweighted estimation method. See also: Laird & Mosteller (1990). Some statistical methods for combining experimental results. International Journal of Technology Assessment in Health Care, 6(1), 5-30.

falls somewhere between β_0 and $\beta_0 + \beta_1$, which neither reflects the correlation for dissertations, nor the correlation for journal articles. Moreover, the estimate of μ will be either closer to β_0 or $\beta_0 + \beta_1$ depending on how many dissertations are included in the meta-analysis relative to the number of journal articles. Finally, the values of θ_i are not randomly scattered around μ as assumed by the random-effects model, but take on only two possible values, namely β_0 and $\beta_0 + \beta_1$. In essence, it is unclear what meaning we should attach to the estimate of μ in such a situation.

As an alternative to a random-effects model analysis, we can actually try to account for the heterogeneity among the θ_i values by modeling the relationship between the outcome measure of interest and the study-level characteristics that we believe exert some influence on the size of the outcome. The process of examining how and to what extent the outcome depends on one or more study-level characteristics is usually called a moderator analysis in the meta-analysis literature. Accordingly, the study-level characteristics examined in such an analysis are typically called moderators.

Potential Moderators

Any study-level variable that may exert a systematic influence on the outcome measure can be considered a potential moderator. For example, when aggregating the results from studies on the effectiveness of a particular treatment (as measured, for example, by a relative improvement rate or a standardized mean difference), we may suspect that treatment length, intensity, or implementation quality influences the outcome in a systematic way. The nature of the comparison or control group may also affect the results (e.g., a treatment may be less effective when the treatment group is compared against a control group that receives some form of standard or alternative care and more effective when the participants assigned to the control group receive no care whatsoever). The type of research design and other methodological characteristics may also be relevant in this context (e.g., whether participants and/or the experimenter were blind with respect to the group assignment can affect the results in predictable ways). Other potential moderators may include characteristics of the subjects (e.g., the effectiveness may be higher in studies excluding patients with comorbidity or depending on the severity of the condition being treated), characteristics of the setting (e.g., results may differ depending on whether a study was conducted in an outpatient or inpatient facility), and the type of measurement instrument used (e.g., self-report and clinician-administered measures may yield different results). When meta-analyzing studies that examine the association between two variables, we may suspect that the strength of the association is influenced, for example, by certain characteristics of the setting, the subjects, and how the variables were operationalized.

Naturally, what moderators are considered relevant will depend on the specific topic at hand and the purpose of the meta-analysis. Moreover, there are often practical limitations to the moderators that can be examined. For one, it is only possible to examine the influence of a particular moderator on the outcome of interest if the studies included in the meta-analysis actually differ with respect to the moderator. For example, if the medication dosage or intervention length was identical in all of the studies included in the meta-analysis, then it is not possible to examine the influence of these moderators on the treatment effectiveness.

The ability to examine certain moderators may also be hampered by a lack of information about the values of the moderators. For example, the effectiveness of a medication or an intervention may be moderated by the degree to which patients adhered to the treatment, but the studies included in the meta-analysis may not report this information (either because of space constraints or because that information is not available to the authors of the studies).

Moderator Analysis via Meta-Regression

Various methods for conducting moderator analyses have been suggested in the literature (e.g., Glass, 1977; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Raudenbush, 1994; Raudenbush & Bryk, 1985; Rosenthal, 1991). The most general and flexible of these approaches is a regression analog for meta-analysis called meta-regression. The specific form of the meta-regression approach described here is based on a mixed-effects model and is essentially a generalization of the random-effects model described earlier.

Specifically, instead of assuming that the θ_i values vary randomly around a single value μ (as assumed by the random-effects model), the mixed-effects model allows the center of the distribution of θ_i values to differ systematically depending on the values of one or more moderator variables. This idea is depicted graphically in Figure 2, which again illustrates a two-step process leading to the observed outcomes.

At the first stage, we now suppose that a distribution of θ_i values is centered at a particular expected value $E(\theta_i)$ that is a linear function of one or more moderators. Therefore, depending on the moderator values, the expected value of the distribution is given by

$$E(\theta_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

where x_{ji} denotes the value of the j^{th} moderator for the i^{th} study, β_j denotes how the expected value of θ_i changes for a one-unit increase in the corresponding x_{ji} value when all other variables are held constant, and β_0 denotes the expected value of θ_i when $x_i = 0$ for $j = 1, \ldots, p$ (note that Figure 2 illustrates the situation where $E(\theta_i)$ is influenced

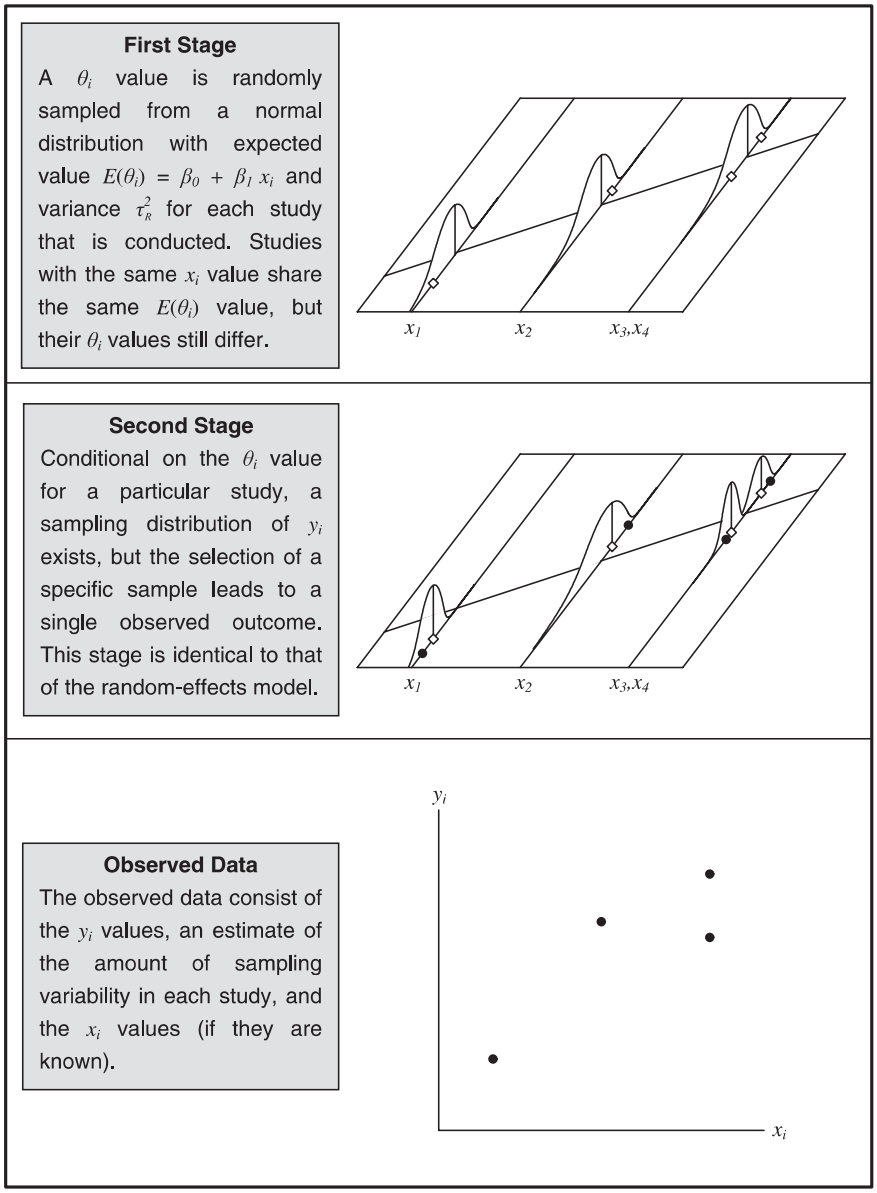

Figure 2. Schematic diagram of the mixed-effects model illustrating the two-step process leading to the observed outcomes.

by only a single continuous moderator like intervention length, but in many cases we would expect that multiple moderators exert an influence on the expected value of θ_i). The variability around $E(\theta_i)$ is assumed to be random, normally distributed, and now denotes "residual heterogeneity," which is the result of other unmeasured factors that introduce additional variability into the θ_i values. In other words, residual heterogeneity is that part of the total variability in the θ_i values that is not accounted for by the moderators included in the model. Therefore, when a study is conducted, the values of the moderators determine $E(\theta_i)$, but a random draw from the distribution around the given $E(\theta_i)$ value yields the actual value of θ_i for that study. Therefore, two studies with the same moderator values share the same value of $E(\theta_i)$, but only by chance will their respective θ_i values coincide.

The second stage of the process leading to the observed outcomes is identical to that discussed under the random-effects model. Specifically, given θ_i for a particular study, an entire distribution of outcomes could potentially be observed, but selecting a single specific sample yields a single observed y_i value for that study. Consequently, the observed y_i value is expected to differ from θ_i because of sampling variability so that two studies that coincidentally share the same θ_i value will still differ with respect to their observed y_i values.

According to the mixed-effects model, we can, therefore, distinguish between three sources of variability in the collection of observed outcomes: the systematic heterogeneity introduced by the influence of moderators on the $E(\theta_i)$ values, the random residual heterogeneity around a particular $E(\theta_i)$ value, and the random sampling variability

around a study-specific θ_i value. Again, only the amount of sampling variability is a function of the sample size within each study.

To actually fit a mixed-effects model to a set of observed outcomes, we first specify what moderators will be included in the model and how the values of the moderators will be coded. Essentially, this process is the same as specifying a regression model in primary research. In fact, the model can encompass the usual extensions for regression models, such as interactions between moderator variables, polynomial moderator terms, and categorical moderators using appropriate dummy coding. A discussion of these topics is beyond the scope of this article, but any standard text on linear models (e.g., Neter, Kutner, Nachtsheim, & Wasserman, 1996) should provide more details.

Having specified a model, we then estimate the amount of residual heterogeneity in the θ_i values (details are given in the appendix). Finally, letting $\hat{\tau}^2_R$ denote the estimated amount of residual heterogeneity, estimates of the regression model parameters and the corresponding standard errors are obtained via weighted least squares (WLS), with weights set equal to $w_i^* = 1/(v_i + \hat{\tau}^2_R)$. In some situations, it may happen that the amount of residual heterogeneity is estimated to be zero. A value of $\hat{\tau}^2_R$ equal to zero suggests that all of the heterogeneity in the θ_i values is accounted for by the moderators included in the model.

Once parameter estimates and corresponding standard errors have been obtained (which we may denote by b_0, . . ., b_p and $SE[b_0]$, . . ., $SE[b_p]$, respectively), it is possible to test whether the influence of a particular moderator on the expected value of θ_i is statistical significant by dividing the respective parameter estimate by its standard error and comparing this ratio against the critical values of a standard normal distribution (i.e., ± 1.96 for $\alpha = .05$, two-tailed). The predicted value of $E(\theta_i)$ for a specific combination of moderator values can also be calculated along with a corresponding 95% confidence interval for $E(\theta_i)$. Equations for fitting the mixed-effects model and carrying out these additional computations are given in the appendix.

In the St. John's wort meta-analysis, we will consider two moderators of treatment effectiveness, namely the treatment intensity and the severity of the condition being treated. If St. John's wort does indeed provide benefits beyond those offered by placebos, we may suspect that the treatment effectiveness increases with the intensity of the treatment. On the other hand, studies conducted with more severely depressed participants may find lower treatment effects, if more severe forms of depression tend to be resistant to treatment with St. John's wort. In fact, these two moderators may interact such that increases in treatment intensity yield corresponding increases in treatment effectiveness for less severe forms of depression, but not for more severe forms of depression. Given these hypotheses, we now must decide how treatment intensity and condition severity will be defined in terms of the available data.

Although treatment intensity could be expressed separately in terms of the treatment duration and the weekly dosage of the *Hypericum* extract used in each study, a new variable was created that is equal to the product of these two moderators. Therefore, for reasons to be outlined later, treatment intensity will be expressed in terms of a single moderator that indicates the total dosage in grams administered during the course of each study. Condition severity will be expressed in terms of the average HRSD score at baseline. The moderator that indicates whether a study was restricted to patients suffering from major depression will not be included in the model, since it overlaps with the baseline moderator to such a great extent as to be virtually redundant (the point-biserial correlation between these two moderators is .84). Finally, since the interaction between total dosage and baseline score will be examined, the product of these two moderators was calculated. To make the interpretation of the results easier, the total dosage and baseline values were centered at 34 and 20, their respectively means rounded to the nearest integer[6].

Therefore, the model of interest stipulates that the expected log relative risk $E(\theta_i)$ is equal to $\beta_0 + \beta_1(D_i - 34) + \beta_2(B_i - 20) + \beta_3(D_i - 34)(B_i - 20)$, where D_i denotes the total dosage and B_i the baseline HRSD score for the i^{th} study. The estimated amount of residual heterogeneity is equal to $\hat{\tau}^2_I = .047$ for this model. Given that the total amount of heterogeneity was estimated to be $\hat{\tau}^2_I = .091$ in the random-effects model, we can conclude that the moderators account for approximately 100(.091 – .047)/.091 = 48% of the heterogeneity in the true log relative rates.

The parameter estimates, corresponding standard errors, the test statistics for the parameters, and the corresponding *p*-values are given in Table 2. Therefore, we see that the expected log relative rate is estimated to be equal to $0.477 - 0.006(D_i - 34) - 0.067(B_i - 20) - 0.002(D_i - 34)(B_i - 20)$. Although just above $\alpha = .05$, the results suggest that St. John's wort is more effective for lower baseline HRSD scores ($z = -1.91$, $p = .06$). On the other hand, the total dosage of St. John's wort administered during the course of a study does not appear to influence the treatment effectiveness ($z = -0.58$, $p = .56$). Finally, the results indicate that the two moderators do not interact ($z = -0.46$, $p = .65$).

The influence of the HRSD value at baseline on the treatment effectiveness becomes more evident if we examine the predicted average effectiveness of St. John's wort for several representative values of this moderator. For example, the expected log relative rate for an average baseline score of 12.5 (the lowest value in the studies considered here) and a total dosage equal to 34 is estimated to be 0.98 with a 95% confidence interval given by (0.38, 1.59). This

[6] Centering the moderator variables changes the interpretation, numerical value, and estimate of β_0, but leaves all other parameters and corresponding estimates unchanged.

Table 2. Results from a mixed-effects model estimating the expected log relative risk as a function of total dosage, baseline HRSD score, and the interaction between these two moderators (the dosage and baseline moderators were centered at their respective means)

Parameter	Parameter estimate	Standard error	z	p
Intercept	0.477	0.0877	5.43	.00
Dosage	–0.006	0.0100	–0.58	.56
Baseline	–0.067	0.0353	–1.91	.06
Dosage × Baseline	–0.002	0.0034	–0.46	.65

corresponds to an average relative rate of exp(0.98) = 2.66 with a 95% confidence interval of (1.46, 4.90). This result suggests that treatment with St. John's wort is, on average, significantly better than placebos for milder forms of depression.

On the other hand, for an average baseline score of 23.6 (the highest value in the studies) and the same total dosage of St. John's wort, the predicted average relative rate is 1.26 with a 95% confidence equal to (0.99, 1.62). Note that the confidence interval now includes the value 1, suggesting that St. John's wort is, on average, no better than placebos. A plot of the observed relative rates against the average HRSD scores at baseline is shown in Figure 3. The size of the points was drawn proportional to the inverse sampling variance, emphasizing the fact that observations are weighted differently in the analysis. The solid line in the figure shows the predicted average effectiveness as a function of baseline HRSD score (holding total dosage constant at 34 grams), while the dashed lines indicate the bounds of the corresponding 95% confidence interval.

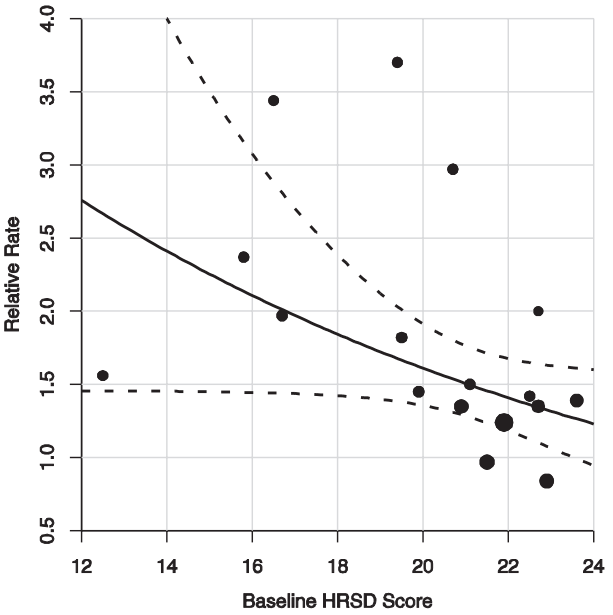

Figure 3. Baseline HRSD score versus the observed relative rate in 17 studies on the effectiveness of St. John's wort for treating depression. The solid line indicates the expected relative rate as predicted from a meta-regression model. The dashed lines indicate the corresponding 95% confidence interval for the expected relative rate.

Interpretation of Moderator Analysis Results

The St. John's wort example demonstrates that it may not be appropriate in some circumstances to view the effect of a treatment or the strength of an association as a single unchanging value. Instead, the treatment effect or association strength may vary systematically as a function of one or more moderator variables. The results from a moderator analysis can have important implications for clinical practice and public policy, as they may suggest under what conditions and for whom a treatment works best or when an association is strongest (Light, 1987; Pillemer & Light, 1980). The proper interpretation of results from a moderator analysis will be considered below by emphasizing some of the limitations inherent to this method.

Correlational Versus Causal Evidence

Moderator analyses are by nature observational studies. In other words, the meta-analyst simply observes, in retrospect, the characteristics of the studies, samples, and methods used and examines whether these features are related to the outcomes in a systematic way. The results from a moderator analysis, therefore, do not provide any evidence of a causal relationship between moderators and outcomes. In particular, it is impossible to rule out the possibility that some unknown third variable influences the observed outcomes and also varies systematically (i.e., is confounded) with the moderator variables of interest, introducing spurious relationships between moderators and observed outcomes that would be absent if the unknown factor could be held constant. Therefore, a moderator analysis should not be regarded as a procedure to test causal relationships, but rather as a method to generate interesting research hypotheses, which can then be examined further via primary research (Cooper, 1998).

Study-Generated Versus Synthesis-Generated Evidence

Despite the limitation that moderator analyses can only provide correlational evidence, moderator analyses allow

meta-analysts to examine relationships that have never been investigated in primary research and, therefore, may provide new insights that could not be obtained from the individual studies (Cooper, 1998). For example, (study-generated) evidence of a (causal) relationship between medication dosage and treatment effectiveness could be obtained by randomly assigning participants to several different dosage conditions within a single study and comparing the results across conditions. However, if the medication dosage was held constant within each study, then information about such a dose-response relationship is lacking. On the other hand, if medication dosage varies across studies, then a moderator analysis can provide (synthesis-generated) evidence about whether the treatment effectiveness is related to the medication dosage.

Ecological Fallacy

When conducting moderator analyses of the type discussed in the present paper, it is important to realize that the unit of analysis is the study and not the individual participant within a study. Accordingly, the moderators we can consider in such analyses consist of study-level characteristics and the relationships we observe, therefore, pertain to relationships between aggregates. However, relationships observed at the aggregate level may not correspond to relationships observed at the individual level. Applying inferences from a higher to a lower level of analysis may lead to the so-called ecological fallacy (Robinson, 1950).

The ecological fallacy can be illustrated with some hypothetical data as shown in Figure 4. Suppose four studies have been conducted to examine the effectiveness of a

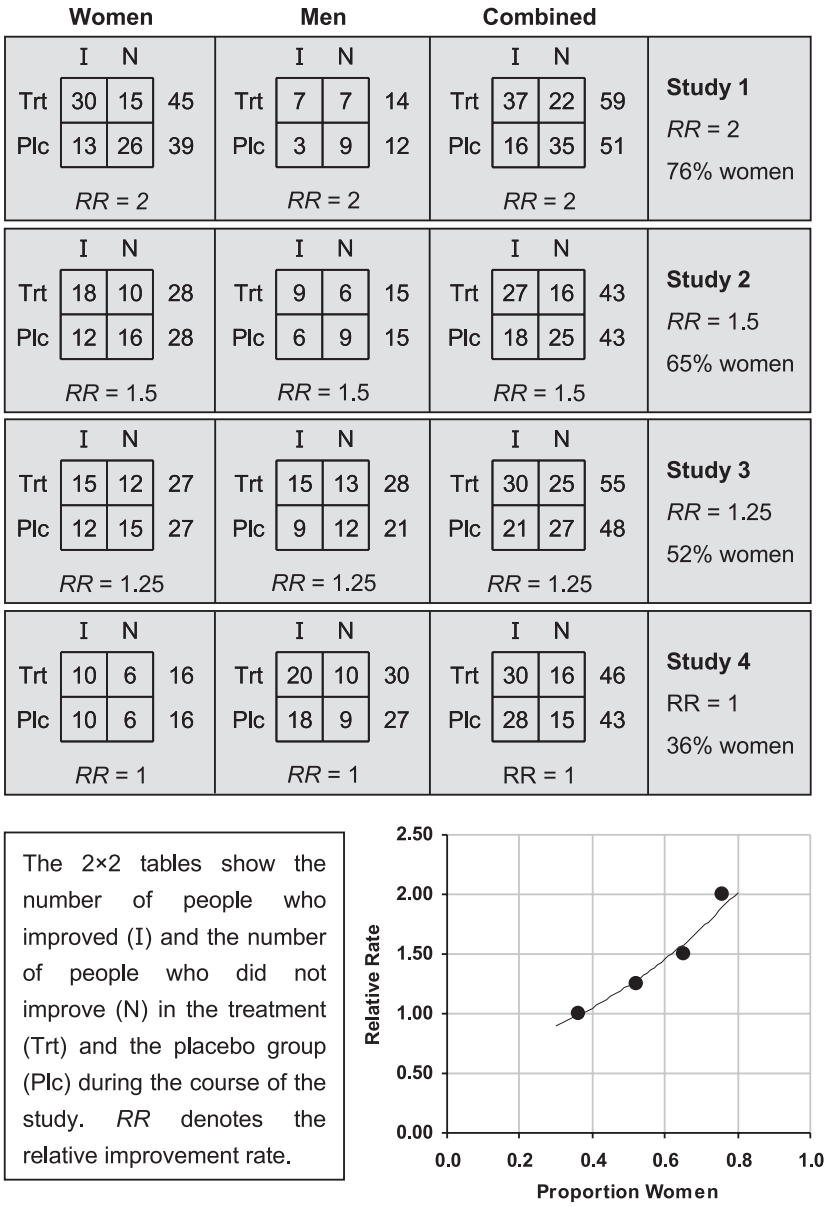

Figure 4. Results from four hypothetical studies on the effectiveness of a treatment illustrating the ecological fallacy.

particular medication. In each of the studies, male and female participants were randomly assigned to receive either the medication or a placebo and the relative improvement rate of the treatment versus the placebo group was recorded after 1 month of treatment. Figure 4 shows the results of each study in the form of three separate 2 × 2 tables, one broken down by gender and one for men and women combined. Note that the relative rate within each study does not depend on gender. Therefore, at the individual level, there is no relationship between treatment effectiveness and gender. Suppose now that a meta-analyst, having access only to the combined results, investigates whether the relative rate depends on the proportion of women in the sample (a moderator that is examined frequently in meta-analyses). The scatterplot in Figure 4 clearly suggests that the relative rate increases systematically with the proportion of women in the sample. In fact, a model that includes the proportion of women in the sample as a moderator yields a highly significant relationship with no residual heterogeneity left (i.e., all of the heterogeneity is accounted for by this moderator). However, interpreting this findings as evidence that the medication is more effective for women than for men would be an example of the ecological fallacy.

The example is admittedly quite extreme, but it nevertheless demonstrates that one must be careful when interpreting the results from a moderator analysis. Moreover, while the example illustrates the case where a relationship found at the aggregate level does not apply to the individual level, it is also possible that a relationship present at the individual level is absent within the aggregated data. A general solution to this problem is to obtain the individual-level data from each study by contacting the primary researchers and to conduct the meta-analysis based on the raw data (Stewart & Tierney, 2002). However, this may not be a feasible option in practice since attempts to obtain the raw data of the individual studies are often unsuccessful (Wicherts, Borsboom, Kats, & Molenaar, 2006).

Practical Issues in Moderator Analyses

There are several practical issues one has to consider when conducting moderator analyses and these are highlighted in this section. Most of these issues actually apply to the analysis of regression-type models in general, but they are discussed here because their importance appears to be underappreciated in the meta-analytic context.

Multiple Testing Problem

The studies included in a meta-analysis often differ with respect to many characteristics, potentially leading to a large number of moderators that could be examined. However, the probability that at least one moderator is found to be statistically significant increases quickly through the accumulation of Type I error probabilities. For example, suppose 10 moderators are tested that are in reality all unrelated to the outcome of interest and each test is conducted with $\alpha = .05$. While each individual moderator test will, therefore, only have a 5% chance of being significant, the probability of obtaining at least one significant result among the 10 tests is equal to $100(1 - .95^{10}) \approx 40\%$. In other words, it is quite likely that at least one moderator is found to be significant by chance alone.

The Bonferroni correction (i.e., dividing the α-value of each test by the number of statistical comparisons to be performed) is among the simplest methods to deal with the multiple testing problem, but it also increases the probability that relevant moderators are going to be overlooked (a more detailed discussion of the multiple testing problem is beyond the scope of the present article, but the interested reader can consult, for example, Shaffer, 1995). Aside from such a purely statistical solution, it is, therefore, advisable to limit the number of moderators tested. Moreover, one should select moderators a priori for inclusion in the model, instead of leaving only those moderators in the model that turn out to be statistically significant.

One method that may be useful to reduce the number of moderators in the model is to combine several moderator variables into some kind of composite score. While the use of a composite score implies a loss of detail regarding the influence of each individual component on the outcome of interest, one can sometimes achieve a drastic reduction in the number of moderator variables with this approach. For example, characteristics that indicate the quality of the studies included in a meta-analysis are frequently examined in moderator analyses. However, instead of including each characteristic by itself in the model, an alternative is to assign an overall quality score to each study that combines the information about the quality characteristics of interest. In fact, numerous quality scales have been developed for this purpose (Moher et al., 1995). A composite score was also used in the St. John's wort example. Treatment intensity was expressed in terms of total medication dosage, combining the dosage and treatment duration information into a single variable.

Finally, a biological, psychological, social, or otherwise plausible reason should exist that could, in theory, explain why and in what direction a particular moderator would be expected to influence the outcome in a systematic manner. Such explanations should be developed before the moderator analysis is actually carried out, since post hoc rationalizations may suffer from hindsight bias.

Sample Size Issues

The importance of keeping the number of moderator variables down to a reasonable level is reinforced by the fact that the number of studies included in meta-analyses is of-

ten relatively small. For example, in 24 meta-analyses from various disciplines, including social, clinical, and organizational psychology, k ranged from 5 to 76 with a mean and median of 24.7 and 18, respectively (Rosenthal & DiMatteo, 2001). Therefore, lessons from research on the statistical properties of regression models when sample sizes are small (e.g., Babyak, 2004) apply with full force in the context of moderator analyses. In particular, when the number of moderators included in the model is large and k is small, then the value of $E(\theta_i)$ for a particular combination of moderator values is going to be estimated so inaccurately as to be essentially useless for practical purposes. Nevertheless, the moderators may still appear to account for a large amount of the heterogeneity even when they are all completely unrelated to the outcome of interest. Therefore, we may be misled to believe that we have actually accounted for much of the heterogeneity when in fact we have simply fitted the noise in the data.

Numerous rules of thumb have been suggested in the context of multiple regression to determine how large the sample size should be relative to the number of predictors in the model in order to obtain robust conclusions (e.g., Green, 1991). While analogous guidelines for moderator analyses are lacking, it needs to be emphasized that such rules are typically overly simplistic and ignore the idiosyncratic nature of different studies. Instead of relying on such crude rules of thumb, an alternative approach is to examine the actual power of moderator tests (Hedges & Pigott, 2004).

Examining Moderators Individually Versus Collectively

An inspection of published meta-analyses reveals that authors typically examine one moderator at a time instead of examining the influence of multiple predictors on the outcome of interest within a single regression model. This practice may, at least in part, be an attempt to avoid the aforementioned problems with small sample sizes. However, since moderator variables are often correlated, the conclusions drawn from a moderator analysis can differ, sometimes dramatically, depending on the approach chosen (Steel & Kammeyer-Mueller, 2002; Viswesvaran & Sanchez, 1998).

For example, consider a series of studies examining the effectiveness of an intervention for treating a particular condition. Assume that about half of the studies were conducted with participants suffering from a more severe form of the condition and for whom the intervention is generally less effective, while the remaining studies were conducted with participants suffering from a less severe form of the condition and for whom the intervention is generally more effective. Now suppose that the intervention length within each study was actually matched to some extent by the experimenters to the condition severity. However, assume that intervention length does not actually have an impact on the treatment effectiveness in reality. Examining the influence of intervention length on the treatment effectiveness without including the condition severity moderator in the model would then lead to the conclusion that short-term interventions are more beneficial. On the other hand, including both moderators simultaneously in the model would reveal that intervention length does not influence the treatment effectiveness.

Prescreening moderators for inclusion in a regression model through a series of individual tests can also lead to problems. On the one hand, certain moderators may appear to be of importance when examined individually, but only because they are correlated (i.e., confounded) with other moderators that actually have an influence on the outcome of interest. Confounding will, therefore, lead to a model that is too complex, which in turn reduces the power to detect actual moderators. On the other hand, certain moderators may not appear to influence the outcome of interest when examined individually, but may play an important role in a model containing other moderators, a condition commonly known as suppression (Horst, 1941). Suppression will, therefore, lead to a model that fails to include relevant moderators.

In general, the practice of examining moderators individually can lead to incorrect conclusions whenever moderators are correlated (Lipsey & Wilson, 2001b). Ideally, models should be constructed to examine the influence of multiple moderators simultaneously. The model coefficients then indicate how much a particular moderator influences the outcome of interest when holding the values of other moderators constant.

Subgrouping and Dichotomization

Another practice frequently seen in the context of meta-analysis is to subgroup the data based on the levels of one or more moderators and/or to examine the relationship between a moderator and the outcome of interest within subsets of studies defined by the levels of other moderators. Going back to the example given in the previous section, one could fit two separate random-effects models for the studies with low and with high severity participants. Furthermore, one could fit separate regressions models examining the relationship between intervention length and treatment effectiveness for each level of the condition severity moderator. While the conclusions reached this way may correspond to those obtained when fitting a model containing both the intervention length and condition severity moderators simultaneously, this approach is less efficient and suffers from another shortcoming. How well the amount of (residual) heterogeneity is estimated depends, in part, on the number of studies included in the analysis (Viechtbauer, 2005). Estimates of (residual) heterogeneity for subsets of studies are, therefore, less precise than a single estimate of (residual)

heterogeneity obtained when analyzing all of the studies together.

Along with the subgrouping approach one can often find cases where naturally continuous moderators (such as intervention length) are dichotomized (e.g., into short and medium/long term interventions). Aside from the loss of information that results from this practice, dichotomization can, on the one hand, reduce the power of moderator tests and, on the other hand, induce relationships where none actually exist (MacCallum, Zhang, Preacher, & Rucker, 2002). Therefore, while dichotomization may simply be a necessary consequence of the lack of detail reported in articles, this practice should be avoided whenever possible.

Relative Importance of Moderators

The fact that moderators are often correlated induces additional problems with respect to the interpretation of the results. When moderators are correlated, then multiple (sets of) moderators may account for the heterogeneity in the θ_i values, which makes it difficult to determine what the most appropriate model should be. Moreover, in models containing multiple correlated predictors, it is difficult to determine how much of the heterogeneity is accounted for by each moderator in the model (Viswesvaran & Sanchez, 1998). Determining the relative importance of predictors and predictor selection are problems that have received extensive attention in the larger context of regression analysis (e.g., Miller, 2002), but relatively little attention has been paid to these issues in meta-analysis. With increased emphasis on the type of model building advocated here, this may change in the future.

Missing Data

The decision to examine one moderator at a time is often a pragmatic one. Reporting practices can vary widely between authors, journals, and publication types and certain information may simply not be available to the meta-analyst based on the study reports. Additional information can sometimes be obtained by contacting the studies' authors, but missing information with respect to the moderator values is usually a common occurrence.

Missing data of this type raise particular problems when examining multiple moderators simultaneously. Specifically, to apply the methods presented in this paper, any study with a missing value for any of the moderator variables included in the model would have to be removed from the analysis (listwise deletion). Given that the number of studies is already quite small in most meta-analyses, deletion of cases with missing data can shrink the size of the dataset to such a degree that it becomes essentially unusable for a moderator analysis. On the other hand, when examining one moderator by itself, only those studies would have to be removed from the analysis with missing data on that particular moderator (pairwise deletion). More data would, therefore, be available for the analysis of individual moderators.

However, even if we assume that all of the moderators are perfectly uncorrelated (so that both approaches could, in principle, yield the same conclusions), simply deleting cases with missing values is a less than ideal practice for dealing with missing data in general (e.g., Schafer & Graham, 2002) as well as in meta-analysis (Pigott, 2001). The unbiasedness of results obtained via case deletion is only assured under certain restrictive assumptions regarding the reason why the data are missing. Essentially, we would have to assume that the studies without missing data are representative of the entire set of studies included in the meta-analysis. On the other hand, distortions are likely to occur when, for example, the reason why the data are missing is related to the value of a moderator itself.

Alternative methods to handle missing data can yield accurate results under less restrictive assumptions regarding the reason why data are missing (e.g., Schafer & Graham, 2002), but how to apply these methods to meta-analysis is still an area of ongoing research (Pigott, 2001). In essence then, the results from a moderator analysis should be interpreted with some reservation when there are extensive holes in the data, regardless of whether moderators are examined individually or in combination.

Testing for Heterogeneity Before a Moderator Analysis

A moderator analysis is only sensible when the θ_i values are actually heterogeneous. The test given by (3) examines whether the amount of variability among the observed outcomes is greater than would be expected based on sampling variability alone. A likely explanation for a significant test statistic is heterogeneity among the θ_i values. Therefore, some meta-analysts will only conduct a moderator analysis when a test for heterogeneity yields a significant result. In that sense, we can regard the testing for heterogeneity as a method that can protect us from making Type I errors in a moderator analysis (since finding that a moderator is significant must be a Type I error by definition when the θ_i values are homogeneous).

However, several studies have shown that heterogeneity tests may have low power to detect heterogeneity when it is indeed present (e.g., Hedges & Pigott, 2001; Sánchez-Meca & Marín-Martínez, 1997; Viechtbauer, 2007). Moreover, given the ubiquity of heterogeneity as suggested by empirical evidence (Field, 2005; Higgins et al., 2003; Lipsey & Wilson, 2001b), one may still conduct a moderator analysis, as long as a model (containing a limited number of moderators) is specified a priori (Hall & Rosenthal, 1991).

Using Standard Regression and Analysis of Variance Models

Some researchers have suggested the use of standard regression and analysis of variance (ANOVA) models to examine the influence of moderators (Glass, 1977; Hall & Rosenthal, 1991). In fact, some of the earliest moderator analyses employed standard regression techniques (e.g., Smith & Glass, 1977). A review of recently published meta-analyses reveals that this practice is still quite common (Steel & Kammeyer-Mueller, 2002). However, given that the amount of sampling variability in outcome measures like the standardized mean difference, the correlation coefficient, or the (log) relative rate is a function of the within-study sample size and given the usual variability in sample sizes across the studies included in a meta-analysis, the homoscedasticity assumption (i.e., a constant error variance for all observations) is essentially guaranteed to be violated in most cases. Therefore, the actual Type I error rate of moderator tests and the actual coverage probability of confidence intervals obtained from such analyses may deviate quite substantially from the nominal values (e.g., Hedges & Olkin, 1985). The techniques described in the present paper are based on weighted least squares methods, which take into consideration the differential amount of sampling variability in the observed outcomes and which generally yield more accurate results (e.g., Steel & Kammeyer-Mueller, 2002).

Software Options for Meta-Regression

Moderator analyses of the type discussed in this paper can be carried out via several general purpose software packages, including SPSS, SAS, R, S-PLUS, and STATA. An alternative option are the software packages MetaWin and Comprehensive Meta-Analysis, which were developed specifically for conducting meta-analyses. A brief overview of these options will be given here.

While SPSS (http://www.spss.com/) does not contain built-in routines to carry out the required computations, a set of macros written by David B. Wilson to accompany the book *Practical Meta-Analysis* (Lipsey & Wilson, 2001a) can be downloaded from the internet (http://mason.gmu.edu/~dwilsonb/ma.html). In particular, the *metareg.sps* macro can be used to fit the meta-regression models considered in the present paper. Users of SAS (http://www.sas.com/) can fit the meta-regression models via the PROC MIXED routine. Tutorials describing this option have been written by Konstantopoulos and Hedges (2004) and Sheu and Suzuki (2001). A meta-regression function for the software packages R (http://www.r-project.org/) and S-PLUS (http://www.insightful.com/) can be downloaded from the author's website (http://www.wvbauer.com/) along with instructions describing its use. User-developed commands for STATA (http://www.stata.com/) also provide the capabilities to conduct moderator analyses by means of meta-regression models (Sharp, 1998; Sterne, Bradburn, & Egger, 2001). Finally, two software packages developed specifically for meta-analytic purposes, namely MetaWin (http://www.metawinsoft.com/) and Comprehensive Meta-Analysis (http://www.metaanalysis.com/), include options for fitting meta-regression models.

Model Extensions and Some Recent Advances

Some extensions of the meta-regression approach presented in this paper and some recent methodological advances are discussed in this final section. The discussion will be brief, but the references given should serve as a starting point to obtain further information.

Dependent Outcomes

Number of studies and number of observed outcomes were used synonymously throughout this article. In essence, it was assumed that each study included in the meta-analysis supplies a single independent estimate of the outcome of interest. In practice one may be able to obtain multiple observed outcomes from a single study. When these estimates are based on a single sample of subjects, then the assumption that the estimates are independent may not be reasonable. The methods presented in the present paper, therefore, have to be extended to take into consideration the amount of covariance between dependent estimates (see, for example, Gleser & Olkin, 1994, for more details).

Estimation of Residual Heterogeneity

Although it may not be apparent at first sight, the estimate of residual heterogeneity in the θ_i values plays a crucial role in the analysis. Small changes in $\hat{\tau}^2_R$ can sometimes yield drastic changes in the conclusions. Several different estimators have been suggested for this parameter (e.g., Raudenbush, 1994; Sidik & Jonkman, 2005; Thompson & Sharp, 1999), but it would be beyond the purposes of the present article to elaborate on their differences. However, the results of Viechtbauer (2005) suggest that the estimator of τ^2_R given in the appendix is approximately unbiased and reasonably efficient.

However, how accurately τ^2_R is actually approximated with this estimator depends on the number of studies, the amount of sampling variability within each study, and the true value of τ^2_R. Given the characteristics of most meta-analyses, $\hat{\tau}^2_R$ may actually be a rather crude estimate and may miss the true value of τ^2_R by a substantial amount.

Therefore, the conclusions from a moderator analysis may not be very robust because of inaccuracies in the estimate of τ^2_R.

One simple yet intuitive method to assess the robustness of the conclusions is to examine how sensitive the results are to slight changes in $\hat{\tau}^2_R$. Essentially, the moderator analysis is repeated multiple times using values of $\hat{\tau}^2_R$ that are gradually shifted away from the actual estimate of $\hat{\tau}^2_R$ (Raudenbush & Bryk, 1985). Conclusions that remain unchanged under such shifts can be regarded as robust.

Improved Moderator Tests

The problem of imprecision in estimates of τ^2_R also affects the results from a moderator analysis on a more general level. In particular, note that the parameters of the regression model given by (7) are estimated via weighted least squares using the weights $w_i^* = 1/(v_i + \hat{\tau}^2_R)$. The moderator analysis actually ignores the fact that $\hat{\tau}^2_R$ is estimated from the data. In fact, not only $\hat{\tau}^2_R$, but also the v_i values are estimates, whose imprecision is ignored. As a result, the actual Type I error probability of moderator tests will often be inflated above the nominal $\alpha = .05$ value. Via simulation methods, it can be shown, for example, that the actual Type I error probability in the St. John's wort meta-analysis for a single moderator is not equal to .05, but actually closer to .08. Although the inflation is not very large in this particular example, there may be cases where the inflation is more severe. Refined methods that account, at least to some extent, for the imprecision in estimates of $\hat{\tau}^2_R$ and v_i have been developed (e.g., Knapp & Hartung, 2003). Using such refined methods, it is possible to bring the true Type I error rate of moderator tests close to the nominal α-value.

Conclusion

Results of meta-analyses have suggested that the outcomes from the included studies are often heterogeneous (Field, 2005; Higgins et al., 2003; Lipsey & Wilson, 2001b). The heterogeneity in the estimated treatment effects or association strengths may be caused by "artificial sources" (Glasziou & Sanders, 2002), such as improper or lack of randomization (potentially leading to noncomparability of the experimental groups at the study beginning), lack of blinding, insufficient follow-up length, or use of measurement instruments with low reliability. Therefore, one of the reasons for the relatively high proportion of meta-analyses with heterogeneous results may be a consequence of ignoring Slavin's (1986) claim that only the most relevant and methodologically sound studies should be included in a meta-analysis. Nevertheless, empirical evidence suggests that at least a quarter of the total amount of heterogeneity in meta-analyses is associated with features of substantive interest (Lipsey & Wilson, 2001b).

Therefore, moderator analyses may be one of the most useful aspects of a meta-analysis, as they may indicate under what circumstances and for whom a particular treatment works best, an association is strongest, or, on a more general level, what study features influence the outcome of interest in a systematic manner (Lau, Ioannidis, & Schmid, 1998; Light, 1987; Lipsey & Wilson, 2001b; Pillemer & Light, 1980). Moderator analyses may even allow us to examine relationships that have never been examined in primary research (Cooper, 1998). However, part of the goal of the present paper was to sensitize the reader to the limitations inherent in this method and to the numerous practical problems one may encounter when conducting a moderator analysis.

The point is not to discourage the use of moderator analyses in general, but to emphasize that we should not ask more of our data than can actually be obtained from them. For example, commonly used approaches to moderator analysis (such as the indiscriminate testing of a large number of individual moderators or the dichotomization of naturally continuous variables) are likely to yield a large number of Type I errors. Given that the results from a moderator analysis is really just the starting point for subsequent research to confirm those findings, much time and effort may be wasted on research that is fruitless from the onset. A model containing a limited number of a priori selected moderators is more likely to yield results that may actually be replicable in future research.

References

Babyak, M.A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66,* 411–421.

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

Chalmers, I., Hedges, L.V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions, 25,* 12–37.

Cooper, H.M. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

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

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

Fleiss, J.L. (1994). Measures of effect size for categorical data. In H.M. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Russell Sage.

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

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

Glasziou, P.P., & Sanders, S.L. (2002). Investigating causes of

heterogeneity in systematic reviews. *Statistics in Medicine, 21,* 1503–1511.

Gleser, L.J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H.M. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York: Russell Sage.

Green, S.B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499–510.

Hall, J.A., & Rosenthal, R. (1991). Testing for moderator variables in meta-analysis: Issues and methods. *Communication Monographs, 58,* 438–448.

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

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

Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L.V., & Pigott, T.D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426–445.

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

Higgins, J.P.T., Thompson, S.G., Deeks, J.J., & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557–560.

Horst, P. (1941). The role of predictor variables which are independent of the criterion. *Social Science Research Council Bulletin, 48,* 431–436.

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

Jackson, G.B. (1980). Methods for integrative reviews. *Review of Educational Research, 50,* 438–460.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22,* 2693–2710.

Konstantopoulos, S., & Hedges, L.V. (2004). Meta-analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 281–297). Thousand Oaks, CA: Sage.

Lau, J., Ioannidis, J.P.A., & Schmid, C.H. (1998). Summing up evidence: One answer is not always enough. *Lancet, 351,* 123–127.

Lecrubier, Y., Clerc, G., Didi, R., & Kieser, M. (1994). Efficacy of St John's wort extract WS 5570 in major depression: A double-blind, placebo-controlled trial. *American Journal of Psychiatry, 159,* 1361–1366.

Lee, W.-L., Bausell, R.B., & Berman, B.M. (2001). The growth of health-related meta-analyses from 1980 to 2000. *Evaluation and the Health Professions, 24,* 327–335.

Light, R.J. (1987). Accumulating evidence from independent studies: What we can win and what we can lose. *Statistics in Medicine, 6,* 221–228.

Light, R.J., & Pillemer, D.B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Linde, K., Berner, M., Egger, M., & Mulrow, C. (2005). St. John's wort for depression: Meta-analysis of randomized controlled trials. *British Journal of Psychiatry, 186,* 99–107.

Lipsey, M.W., & Wilson, D.B. (2001a). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Lipsey, M.W., & Wilson, D.B. (2001b). The way in which intervention studies have "personality" and why it is so important to meta-analysis. *Evaluation and the Health Professions, 24,* 236–254.

MacCallum, R.C., Zhang, S., Preacher, K.J., & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40.

McDaniel, M.A., Whetzel, D.L., Schmidt, F.L., & Maurer, S.D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

Meadows, A.J. (1998). *Communication research*. San Diego: Academic Press.

Miller, A. (2002). *Subset selection in regression* (2nd ed.). London: Chapman & Hall.

Montgomery, S.A., Hübner, W.D., & Grigoleit, H.G. (2000). Efficacy and tolerability of St John's wort extract compared with placebo in patients with a mild to moderate depressive disorder. *Phytomedicine, 7*(suppl. 2), 107.

Moher, D., Jadad, A.R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. (1995). Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. *Controlled Clinical Trials, 16,* 62–73.

Neter, J., Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Irwin.

Pigott, T.D. (2001). Missing predictors in models of effect size. *Evaluation and the Health Professions, 24,* 277–307.

Pillemer, D.B., & Light, R.J. (1980). Benefiting from variation in study outcomes. In R. Rosenthal (Ed.), *New directions for methodology of social and behavioral science: Quantitative assessment of research domains* (Vol. 5, pp. 1–12). San Francisco: Jossey-Bass.

Raudenbush, S.W. (1994). Random effects models. In H.M. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage.

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

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.

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

Rosenthal, R. (1994). Parametric measures of effect size. In H.M. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage.

Rosenthal, R., & DiMatteo, M.R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52,* 59–82.

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

Sánchez-Meca, J., & Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type I error. *Quality and Quantity, 31,* 385–399.

Schafer, J.L., & Graham, W.K. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

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

Shaffer, J.P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46,* 561–584.

Sharp, S. (1998). Meta-analysis regression. *Stata Technical Bulletin, 42*, 16–24.

Sheu, C.-F., & Suzuki, S. (2001). Meta-analysis using linear mixed models. *Behavior Research Methods, Instruments, and Computers, 33,* 102–107.

Sidik, K., & Jonkman, J.N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Applied Statistics, 54,* 367–384.

Slavin, R.E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15,* 5–11.

Smith, M.L. (1980). Publication bias and meta-analysis. *Evaluation in Education, 4,* 22–24.

Smith, M.L., & Glass, G.V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Steel, P.D., & Kammeyer-Mueller, J.D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology, 87,* 96–111.

Sterne, J.A.C., Bradburn, M.J., & Egger, M. (2001). Meta-analysis in Stata. In M. Egger, G.D. Smith, & D.G. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (pp. 347–369). London: BMJ Books.

Stewart, L.A., & Tierney, J.F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation and the Health Professions, 25,* 76–97.

Thompson, S.G., & Sharp, S.J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine, 18,* 2693–2708.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30,* 261–293.

Viechtbauer, W. (2007). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology, 60,* 29–60.

Viswesvaran, C., & Sanchez, J.I. (1998). Moderator search in meta-analysis: A review and cautionary note on existing approaches. *Educational and Psychological Measurement, 58,* 77–87.

Wicherts, J.M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61,* 726–728.

Wolfgang Viechtbauer

Department of Methodology and Statistics
University of Maastricht
P.O. Box 616
NL-6200 MD Maastricht
The Netherlands
Tel. +31 43 388-2277
Fax +31 43 361-8388
E-mail wvb@wvbauer.com

Appendix

Equations for conducting a moderator analysis via the meta-regression approach described in the paper are given in this appendix. Writing out the equations in algebraic notation is extremely cumbersome when considering more than one moderator variable simultaneously. The equations are, therefore, given here in matrix notation.

Let $\boldsymbol{y} = [y_1 \ldots y_k]'$ denote the $(k \times 1)$ vector of observed outcomes. Moreover, let $\boldsymbol{X}$ denote the $(k \times (p + 1)$ matrix containing the values of the p moderators, with the first column in $\boldsymbol{X}$ consisting entirely of 1's corresponding to the intercept. Finally, let $\boldsymbol{W}$ denote a $(k \times k)$ diagonal matrix, with elements equal to $w_i = 1/v_i$. An estimate of the amount of residual heterogeneity can then be obtained with

$$\hat{\tau}^2_R = \frac{\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y} - (k-p-1)}{tr[\boldsymbol{P}]},$$

where $\boldsymbol{P} = \boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}$ (e.g., Raudenbush, 1994). This estimator is, in fact, a generalization of the DerSimonian-Laird estimator of the total amount of heterogeneity given by (4). A negative estimate of τ^2_R is truncated to zero.

Setting the diagonal elements of the $\boldsymbol{W}$ matrix equal to $w_i^* = 1/(v_i + \hat{\tau}^2_R)$, the parameter estimates are then obtained with

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}.$$

The variance-covariance matrix of the parameter estimates is obtained with

$$\hat{\boldsymbol{\Sigma}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}.$$

Taking the square root of the diagonal elements of $\hat{\boldsymbol{\Sigma}}$ yields the standard errors of the parameter estimates. The influence of a moderator variable on $E(\theta_i)$ can then be tested by dividing the respective parameter estimate by its standard error, which can be compared against the critical values of a standard normal distribution. For values x_{1i} through x_{pi} of the moderator variables, the predicted value of $E(\theta_i)$ is equal to $\boldsymbol{x}_i\boldsymbol{b}$, where $\boldsymbol{x}_i = [1\ x_{1i} \ldots x_{pi}]$. A corresponding 95% confidence interval is given by

$$\boldsymbol{x}_i\boldsymbol{b} \pm 1.96\sqrt{\boldsymbol{x}_i\hat{\boldsymbol{\Sigma}}\boldsymbol{x}_i'}.$$