Research Article

Research Synthesis Methods

Received 1 December 2009, Revised 19 August 2010, Accepted 25 August 2010 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.12

A basic introduction to fixed-effect and random-effects models for meta-analysis

Michael Borenstein[a]*†, Larry V. Hedges[b], Julian P.T. Higgins[c] and Hannah R. Rothstein[d]

There are two popular statistical models for meta-analysis, the fixed-effect model and the random-effects model. The fact that these two models employ similar sets of formulas to compute statistics, and sometimes yield similar estimates for the various parameters, may lead people to believe that the models are interchangeable. In fact, though, the models represent fundamentally different assumptions about the data. The selection of the appropriate model is important to ensure that the various statistics are estimated correctly. Additionally, and more fundamentally, the model serves to place the analysis in context. It provides a framework for the goals of the analysis as well as for the interpretation of the statistics.

In this paper we explain the key assumptions of each model, and then outline the differences between the models. We conclude with a discussion of factors to consider when choosing between the two models. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: meta-analysis; fixed-effect; random-effects; statistical models; research synthesis; systematic reviews

Introduction

There are two popular statistical models for meta-analysis, the fixed-effect model and the random-effects model. Under the fixed-effect model we assume that there is one true effect size that underlies all the studies in the analysis, and that all differences in observed effects are due to sampling error. While we follow the practice of calling this a fixed-effect model, a more descriptive term would be a common-effect model. In either case, we use the singular (*effect*) since there is only one true effect.

By contrast, under the random-effects model we allow the true effect sizes to differ—it is possible that all studies share a common effect size, but it is also possible that the effect size varies from study to study. For example, the effect size might be higher (or lower) in studies where the participants are older, or more educated, or healthier than in other studies, or when a more intensive variant of an intervention is used. Because studies will differ in the mixes of participants and in the implementations of interventions, among other reasons, there may be different effect sizes underlying different studies. If it were possible to perform an infinite number of studies (based on the inclusion criteria for the review), the true effect sizes for these studies would be distributed about some mean. In a random-effects meta-analysis model, the effect sizes in the studies that actually were performed are assumed to represent a random sample from a particular distribution of these effect sizes (hence the term *random effects*). Here, we use the plural (*effects*) since there is an array of true effects.

The selection of the model is critically important. In addition to affecting the computations, the model helps to define the goals of the analysis and the interpretation of the statistics. In this paper we explain the similarities and differences between the models and discuss how to select an appropriate model for a given analysis.

Motivating example

For illustrative purposes, we use fictional scenarios in which the goal is to estimate the mean score on a science aptitude test. This example is a bit unusual, in that the effect size is a simple mean, whereas most meta-analyses employ an effect size that

[a] *Biostat, Inc., Englewood, NJ, U.S.A.*
[b] *Department of Statistics, Northwestern University, Evanston, IL, U.S.A.*
[c] *MRC Biostatistics Unit, Cambridge, U.K.*
[d] *Management Department, Baruch College—City University of New York, NY, U.S.A.*
* *Correspondence to: Michael Borenstein, Biostat, Inc., Englewood, NJ, U.S.A.*
† *E-mail: MichaelB@PowerAndPrecision.com*

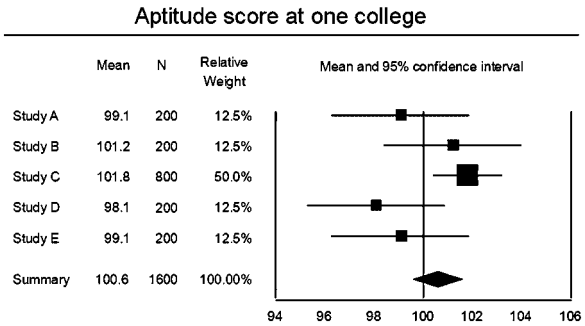

Figure 1. Example of a fixed-effect analysis.

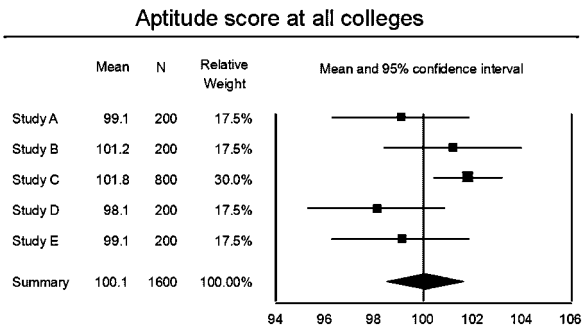

Figure 2. Example of a random-effects analysis.

measures the impact of an intervention or the strength of a relationship. However, the same procedures apply in all cases, and the selected example will allow for a simpler presentation of the relevant issues.

Fixed-effect example

The defining feature of the fixed-effect model is that all studies in the analysis share a common effect size. Suppose that we want to estimate the mean aptitude score for freshmen at a specific college. Suppose further that the true mean at this college is 100, with a standard deviation of 20 points and a variance of 400. We generate a list of 1600 freshmen (selected at random) from that college. The testing facility cannot accommodate all these students at one sitting, and so we divide the names into five groups, each of which is considered a separate study. In studies A, B, D, and E, the sample size is 200, whereas in study C the sample size is 800. If we assume that the assignment to one group or another has no impact on the score, then all five studies share a common (true) effect size, and the fixed-effect model applies.

The analysis based on a fixed-effect model is shown in Figure 1. The effect size and confidence interval for each study appear on a separate row. The summary effect and its confidence interval are displayed at the bottom.

Random-effects example

The defining feature of the random-effects model is that there is *a distribution* of true effect sizes, and our goal is to estimate the mean of this distribution. Suppose that we want to estimate the mean score for freshmen at *any* college in California. Suppose further that the true mean across all of these colleges is 100 and that within any college the scores are distributed with a standard deviation of 20 points and variance of 400. First, we select five colleges at random. Then, we sample students at random from each of these colleges, using a sample size of 200 for colleges A, B, D, and E, and of 800 for college C. We will be using the mean of these colleges to estimate the mean at all colleges. Since it is possible (indeed, likely) that the true mean differs from college to college, the fixed-effect model no longer applies, and a random-effects model is more plausible.

The analysis based on a random-effects model is shown in Figure 2. The effect size and confidence interval for each study appear on a separate row. The summary effect and its confidence interval are displayed at the bottom.

Note that we have deliberately (if somewhat artificially) used the same data for the fixed-effect and random-effects models to highlight how the underlying model affects the results. The following differences are apparent. First, the confidence interval for the summary effect is wider under the random-effects model. Second, the study weights are more similar under the random-effects model (large studies lose influence while small studies gain influence). For example, compare the size of the boxes for the large study (C) vs a small study (B) under the two models. Third, the estimate of the effect size differs under the two models.

The two analyses sometimes produce exactly the same results, but often their results will differ. When this happens, either model can yield the higher estimate of the effect size. Furthermore, under the random-effects model the confidence interval will

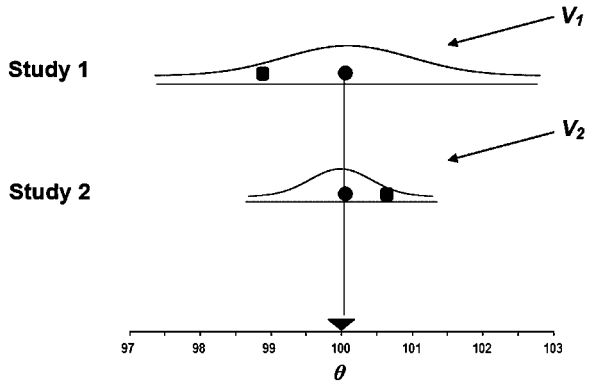

Figure 3. Schematic of the fixed-effect model.

always be wider and the weights will always be more similar to each other than under the fixed-effect model. The reason for these different results will become clear as we explore the differences between the two models.

Deriving the combined effect

Sources of variance in meta-analysis

We use the term *variance* in five ways, to refer to (a) the square of the standard deviation of scores in the population (the *population variance*), (b) the square of the standard deviation of true effects across studies (the *between-studies variance*), (c) the square of the standard error of estimation within a study (the *within-study error variance*), (d) the square of the standard error of estimation in the context of the meta-analysis model (the *overall study error variance*) and (e) the square of the standard error of estimation of the combined result (the *meta-analysis error variance*). Note that (a) and (b) are properties only of the distributions of scores within and across populations, and as such do not depend on the sample size. By contrast, (c), (d) and (e) depend also on the sizes of the samples, and to clarify this we refer to them all as error variances.

The majority of meta-analyses assign weights to each study based on the inverse of the overall study error variance (that is, 1/variance), and that is the approach we shall pursue in this paper. This provides a generic approach to meta-analysis that can be used to combine estimates of a large variety of metrics, including standardized mean differences, means differences, correlation coefficients, regression coefficients, odds ratios and simple means or proportions. Studies with a precise estimate of the population effect size (a low variance) are assigned more weight, while studies with a less precise estimate of the population effect size (a high variance) are assigned less weight.

This scheme is used for both the fixed-effect and random-effects models. Where the models differ is in what we mean by a precise estimate, or (more correctly) how we define the overall study error variance. As outlined above, under the fixed-effect model there is one level of sampling (we sample subjects within the college), and therefore one source of variance. By contrast, under the random-effects model there are two levels of sampling (we sample colleges from the population of colleges and then students within each college), and therefore two sources of variance.

The fixed-effect model is depicted in Figure 3. All studies share a common (true) effect size, which is a score of 100. This mean is represented by the symbol θ (theta) at the bottom of the figure.

The figure shows two studies drawn from this population. For each study, the *true* score is represented by a filled circle. The circle for each study falls at the common θ (100), since all studies are assumed to share the same effect size. The *observed* mean for each study is represented by a filled square, which differs from the true mean because of estimation error.

The observed mean for any study *i* is given by

$$Y_i = \theta + \varepsilon_i, \tag{1}$$

where ε_i is the difference between the common true mean and the observed mean for study *i*. It follows that there is only one source of variation, the estimation error ε_i. In a fixed-effect meta-analysis, the overall study error variance is equal to this within-study error variance.

For each study, we have superimposed a normal curve on the true score. This curve is based on the within-study error variance and shows the range within which the observed mean score is likely to fall. This variance is given by

$$V_i = \frac{\sigma^2}{n}. \tag{2}$$

In words, the error variance of a sample mean depends on two factors. One is the variance of the individual observations in the population (σ^2)—the sample means will be more precise if the scores are clustered in a narrow range. The other is the size of the sample (n)—a larger sample yields a more precise estimate of the population mean.

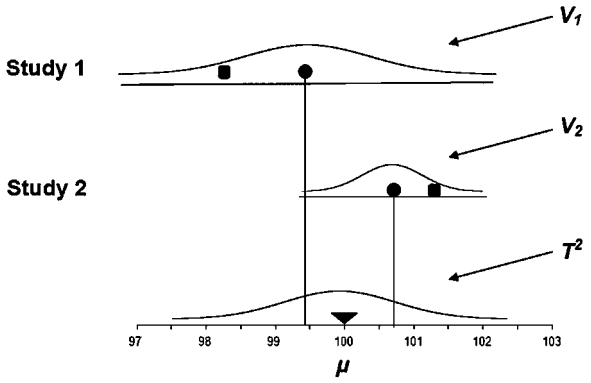

Figure 4. Schematic of the random-effects model.

The population variance for both studies is assumed to be 400. For Study 1 the sample size is 200, which yields a within-study error variance of 2.000 and standard error of 1.414. For Study 2 the sample size is 800, which yields an error variance of 0.500 and standard error of 0.707. With an infinitely large sample size, the width of the normal curve will approach zero and the observed mean for the study will approach the true mean for that study (which is also the common mean).

The random-effects model is depicted in Figure 4. The mean of all schools (100) is represented at the bottom by the symbol μ (mu). We use μ here since this value is the mean of a distribution, as opposed to the θ that we used to represent a common effect size in Figure 3.

The observed mean Y for any study i is given by

$$Y_i = \mu + \xi_i + \varepsilon_i \tag{3}$$

where ξ_i is the difference between the grand mean (μ) and the true mean (θ_i) for study i ($\xi_i = \theta_i - \mu$) and ε_i is the difference between the true mean for study i (θ_i) and the observed mean (Y_i) for study i ($\varepsilon_i = Y_i - \theta_i$) It follows that there are two sources of variation, the variance of ξ and the variance of ε. As we shall see, the overall study error variance in a random-effects meta-analysis includes these two components.

We have superimposed a normal curve above μ, representing the distribution of true means across all colleges. The curve extends roughly from 97.5 to 102.5. This indicates that if we were to perform the study at every school in California, and test all freshmen at each school (so that we knew the true score at each school), the mean scores would be distributed as in this curve.

The standard deviation of the distribution depicted by this curve is called τ (tau) and the variance is called τ^2, analogous to σ (sigma) and σ^2 in a primary study. The sample estimates of τ and τ^2 are denoted by T and T^2, analogous to S and S^2 in a primary study. The majority of the curve extends two standard deviations on either side of the mean, so any study sampled at random from this distribution will usually have a true mean somewhere in this range.

The figure shows two studies drawn from this population, one with a true score (θ_1) of 99.400 and one with a true score (θ_2) of 100.700. For each study, the *true* score is represented by a filled circle. The location of the circle is no longer fixed at 100, but now varies from study to study since the true scores have been sampled from the distribution at the bottom. *This is the key difference between the two models.*

The *observed* score for each study is represented by a filled square, which invariably differs from the true score *for that study* because of estimation error. The population variance for both studies is assumed to be 400. For Study 1 the sample size is 200, which yields a sample variance of 2.0 and standard error of 1.414. For Study 2 the sample size is 800, which yields a sample variance of 0.500 and standard error of 0.707. If the sample size in any study would approach infinity, the width of the normal curve *for that study* would approach zero, and the observed mean for that study would approach the true mean *for that study*. For Study 1 it would approach 99.400, and for Study 2 it would approach 100.700.

Inverse variance weights

We have addressed the issue of variance in some detail because the variance is central to all computations and to the difference between the models. For the purpose of computing a summary effect we will want to assign more weight to studies that yield a more precise estimate of that effect.

Under the fixed-effect model, the overall study error variance for each study's observed mean (Y) about the common mean (θ) is simply the within-study error variance. Thus, in Figure 3 the variance of Y_1 about θ is V_1, and the variance of Y_2 about θ is V_2. More generally, the variance of any study's observed score Y_i about θ is V_i. Therefore, the weight assigned to each study under the inverse variance scheme is simply

$$W_i = \frac{1}{V_i}. \tag{4}$$

By contrast, under the random-effects model, because we use the study means to estimate μ (the grand mean), there are two sources of variance, and therefore two components to the overall study error variance. First, the observed value Y for any study (i) differs from that study's true value (θ_i) because of within-study error variance, V_i. Second, the *true* value (θ_i) for each study differs from μ because of between-study variance. In Figure 4, the variance of Y_1 about θ_1 is V_1 and the variance of θ_1 about μ is τ^2, so the variance of Y_1 about μ is V_1 plus τ^2. Similarly, the variance of Y_2 about θ_2 is V_2 and the variance of θ_2 about μ is τ^2, so the variance of Y_2 about μ is V_2 plus τ^2. More generally, the variance of any observed score Y_i about μ is V_i plus τ^2. Therefore, the weight assigned to each study under the inverse variance scheme is

$$W_i = \frac{1}{V_i + T^2}. \tag{5}$$

Note that the within-study error variance, V_i, is unique to each study, but the between-study variance T^2 is common to all studies.

Note also that formulas (4) and (5) are identical except for the inclusion of T^2 in the latter. If we are using the random-effects model and τ^2 is estimated at zero (i.e. all studies appear to share the same true effect size), then the random-effects model collapses to the fixed-effect model and the two formulas yield the identical result. This is consistent with the idea that the random-effects model allows that the effect size may (but also *may not*) vary from study to study. For details of computing T^2, see Box 1.

Computing the combined effect

Once we have assigned a weight to each study, the combined effect is given by the weighted mean across all studies. If W_i is the weight assigned to study i, Y_i is the observed effect size in study i, and k is the number of studies in the analysis, then the weighted mean (M) is computed as:

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}. \tag{6}$$

Some texts use W to refer to the fixed-effect weight and W^* to refer to the random-effects weight. Here, we have used W to refer to both in order to highlight the parallel between the two models. In both cases, the weight is based on the inverse of the overall study error variance, and the only difference between the models is in how that variance is defined.

Addressing uncertainty in the combined effect

Having computed an estimate of the combined effect, we must also compute the precision of that estimate. We do so by working with the study weights.

Recall that for a single study the weight is the inverse of the overall study error variance,

$$W = \frac{1}{V}. \tag{7}$$

It follows that the overall study error variance is also the inverse of the weight,

$$V = \frac{1}{W}. \tag{8}$$

This holds true also *across* studies (that is, for the *sum* of the weights). Concretely, the *meta-analysis* error variance, V_M, i.e. that of the combined effect M, is given by

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i}, \tag{9}$$

where the study weights, W_i, are summed across studies 1 to k. If we think of the study weight as the *information* contributed by a study, then the sum of weights is the total information available to estimate M, and the higher this value, the lower the variance. As was true for the computation of the combined effect, this formula is the same for both the fixed-effect and the random-effects models (see Appendix A).

The standard error of the combined effect M, then, is simply the square root of the meta-analysis variance,

$$\mathrm{SE}_M = \sqrt{V_M}. \tag{10}$$

Confidence interval

The 95% lower and upper limits for the summary effect are estimated as

$$\mathrm{LL}_M = M - 1.96 \times \mathrm{SE}_M \tag{11}$$

and

$$\mathrm{UL}_M = M + 1.96 \times \mathrm{SE}_M, \tag{12}$$

where the standard error of the combined effect is simply the square root of the variance.

Confidence intervals for other confidence levels (e.g. 90% or 99%) are computed by replacing 1.96 with the appropriate standard normal critical value (e.g. 1.64 for 90% confidence intervals).

Testing the null hypothesis

Under the fixed-effect model we can test the null hypothesis that the *common* true effect size (θ) is a specific value X_0, where X_0 is usually zero (though it would be something else in the motivating example). Under the random-effects model we can test the null hypothesis that the *mean* true effect size (μ) is a specific value (usually zero). In either case, the test statistic (Z) is given by:

$$Z = \frac{M - X_0}{\mathrm{SE}_M}. \tag{13}$$

For a one-tailed test the p-value is given by

$$p = 1 - \Phi(\pm|Z|), \tag{14}$$

where we choose '+' if the difference is in the expected direction and '–' otherwise, and for a two-tailed test by

$$p = 2[1 - (\Phi(|Z|))] \tag{15}$$

where $\Phi(Z)$ is the standard normal cumulative distribution. This function is tabled in many introductory statistics books and is implemented in Microsoft Excel, for example, as the function NORMSDIST(Z).

Understanding uncertainty in the combined effect

Intuitively, the variance of the combined effect size should be smaller under the fixed-effect model (where there is only one source of variance) than under the random-effects model (where there is an additional source of variance). In fact, this logic is built into the formula. Under the random-effects model the overall error variance for each study will be larger (since it includes the fixed-effect variance *plus* T^2), and the weight smaller, than under the fixed-effect model.

If the weight for *each* study is smaller under the random-effects model, then the sum of the weights will be smaller, so the variance and standard error of the combined effect will be larger, and the confidence interval will be wider. In addition, because the standard error is larger, for any given effect size the Z-value will be smaller and the p-value will be closer to 1.0.

The key issue, however, is not that the standard error is larger under the random-effects model, but that it is larger because it is the standard error of an estimate of *a different parameter*, and this estimate has two sources of variation. Under the fixed-effect model we are estimating the mean in one population and the standard error reflects the precision of *that* estimate. Under the random-effects model we are estimating the mean of many populations and the standard error reflects the precision of *that* estimate.

Beyond the broad notion that the standard error is always at least as large (and often larger) under the random-effects model as compared with the fixed-effect model, it is helpful to understand what factors influence the precision of this estimate and how this is different under the two models.

In a primary study the precision of the estimate of the population mean depends on two factors. One is how widely the scores are dispersed in the population, represented by the standard deviation. If the scores are dispersed widely (a high standard deviation) the estimate will tend to be less precise, whereas if the scores are clustered in a narrow range (a small standard deviation) the estimate will tend to be more precise. The other factor is the size of the sample. As the sample size increases, the sample mean will tend to approach the true mean. Concretely, the within-study error variance of the sample mean is given by

$$V = \frac{\sigma^2}{n}, \tag{16}$$

where the numerator captures the first factor and the denominator captures the second.

The mechanism for the fixed-effect meta-analysis is essentially the same as it is for the primary study. Indeed, since all studies in the analysis are estimating the same value, for the purpose of computing the meta-analysis error variance we can think of all

the information as coming from one large study. Concretely, if we assume that every study has the same population variance, σ^2, then the variance of the combined effect is given by

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i} = \frac{1}{\sum_{i=1}^{k} \frac{n_i}{\sigma^2}} = \frac{\sigma^2}{\sum_{i=1}^{k} n_i}. \quad (17)$$

where the number of subjects in each study (n_i) is summed across the k studies. If we also assume that every study has the same sample size, n, this can be expressed as:

$$V_M = \frac{\sigma^2}{k \times n}. \quad (18)$$

In either case, the variance of the combined effect is simply the population variance divided by the total sample size.

Importantly, because all studies are estimating the *same* value, it does not matter how the subjects are distributed across the studies. The variance of the combined effect will be the same regardless of whether we have 10 studies with 100 persons in each, 5 studies with 200 in each, or 1 study with 1000 persons, because the cumulative sample size is the same in all cases. This last example is not strictly a meta-analysis, but highlights the idea of a common effect size—we can think of it as one study divided into 10, 5, or 1 parts. In fact, when there is only one study, (16), (17) and (18) are identical.

When we turn to the random-effects model, we are faced with a different situation. Now, there are two sources of variance. Between-studies variance refers to the fact that the true mean of our k studies is not the same as the true mean of all studies that could have been performed. Within-study error variance refers to the fact that the observed mean for each study is not the same as the true mean for that study.

Importantly, the two sources of variance function independently of each other. Concretely, if we assume again that every study has the same population variance and the same sample size, then the variance of the combined effect is given by:

$$V_M = \frac{\sigma^2}{k \times n} + \frac{T^2}{k}. \quad (19)$$

The first term, which reflects within-study error, is identical to that for the fixed-effect model. With a large enough sample size, this term will approach zero. By contrast, the second term (which reflects the between-studies variance) will only approach zero as the number of *studies* approaches infinity.

To see the origin of the second term, we can isolate the impact of between-studies variance on the accuracy of the combined effect by imagining that the sample size within each study approaches infinity. It follows that the true effect for each study is known with essentially no estimation error. The precision with which we estimate the mean effect across all studies then depends on two factors. One is how widely the true study means are dispersed (if they cluster in a narrow range, the precision will be higher). The other is how many studies are being used to make the estimate (with more studies the precision will be higher). Concretely, in this case the variance of the combined effect is given by:

$$V_M = \frac{T^2}{k}. \quad (20)$$

Note that this is analogous to (16), the formula for the variance of the sample mean in a primary study. In the primary study we were dealing with one mean and the observations were subjects. Here we are dealing with a grand mean and the observations are studies. But the logic is the same and the formulas have the same structure.

The consequence of this second term in the variance of the combined effect (19) is that, if we have only a few studies, there is a limit to the precision with which we can estimate the grand mean, no matter how large the sample size in these studies. Only as k approaches infinity will the impact of between-study variation approach zero. These formulas do not apply exactly in practice, but the conceptual argument does. Namely, increasing the sample size within studies is not sufficient to reduce the standard error beyond a certain point (where that point is determined by T^2 and k). If there is only a small number of studies, then the standard error could still be substantial even if the total n is in the tens of thousands or higher.

A point implicit in the above should be stated clearly. Some people assume that a very large study always represents the *gold standard* and typically yields a 'better' estimate than a meta-analysis of smaller studies. The large-scale study may yield a very precise estimate of the effect *in one study*, but says nothing about the effect in other studies nor (it follows) of the mean effect across a universe of studies. If the true effect varies from study to study, and if our goal is to estimate the overall mean, then a meta-analysis that incorporates many moderately sized studies may well yield a more accurate estimate of the mean effect than the single large-scale study. Additionally, the meta-analysis will address the dispersion in effects, which the single study cannot.

Differences between fixed-effect and random-effects meta-analyses

Factors that affect the estimate of the combined effect

While many researchers find it intuitive that the random-effects model yields a less *precise* estimate of the combined effect than does the fixed-effect model, some find it surprising that the random-effects model also yields a *different* estimate of the combined

effect itself. Indeed, while the fixed-effect model is estimating a common mean and the random-effects model is estimating the grand mean, one might expect the two estimates, based on the same set of study means, to be the same.

In fact, though, the estimate of the combined effect is almost always different under the two models (provided, of course, that T^2 is not zero). The reason is that the combined mean is computed as the weighted mean of the effect size in each study, and the weights are different under the two models.

It is instructive to consider how the weights differ under the two models, and how this will affect the estimate of the combined effect. We present an intuitive explanation and then show how this is implemented in the computations that were presented earlier.

Under the fixed-effect model we assume that the true effect size for all studies is identical and the only reason that the effect size varies between studies is the within-studies estimation error (error in estimating the effect size). Therefore, when assigning weights to the different studies we can largely ignore the information in the smaller studies since we have better information about *the same* effect size in the larger studies.

This is implemented in the weighting scheme. Under the fixed-effect model the weights are based solely on the within-study variances. In the running example, a study with 800 students has one-fourth the variance of a study with 200 people, and is assigned four times as much weight.

By contrast, under the random-effects model the goal is not to estimate one true effect, but to estimate the mean of a distribution of effects. Since each study provides information about *a different* effect size, we want to be sure that all these effect sizes are represented in the summary estimate. This means that we cannot discount a small study by giving it a very small weight (the way we would in a fixed-effect analysis). The estimate provided by that study may be imprecise, but it is information about an effect that no other study has estimated. By the same logic we cannot give too much weight to a very large study (the way we might in a fixed-effect analysis). Our goal is to estimate the mean effect in a range of studies and we do not want that overall estimate to be overly influenced by any one of them.

This is implemented in the weighting scheme when the weights are based on the within-study variance plus a constant (T^2, the between-study variance). Because T^2 is a constant, it reduces the relative differences among the weights, which means that the relative weights assigned to each study are more balanced under the random-effects model than they are under the fixed-effect model.

Concretely, if we move from fixed-effect weights to random-effects weights, large studies lose influence and small studies gain influence. If the larger studies have high effects, then the combined effect will tend to be higher under the fixed-effect model (where these studies have more influence), and lower under the random-effects model (where these studies have less influence). If the smaller studies have high effects, then the combined effect will tend to be lower under the fixed-effect model (where these studies have less influence) and higher under the random-effects model (where these studies have more influence).

While study weights are always more similar to each other when using random-effects weights as compared with fixed-effect weights, the extent of similarity will depend on the ratio of within-study error variance to between-studies variance. At one extreme, if there is substantial within-study variance and only trivial between-studies variance, the weights will be driven primarily by sample size for each study. At the other extreme, if there is minimal within-study variance and substantial between-studies variance, the weights will be almost identical for all studies. Usually, the situation falls somewhere between the two extremes.

Note that the impact of T^2 on the relative weights *does not* depend on how many studies are included in the analysis, because the issue here is how much uncertainty T^2 adds to the effect size estimate in a specific study. By contrast, the precision of the combined effect (above) *does* depend on the number of studies because the issue there is how much information each study adds to this summary.

The context

To this point we have focused on specific elements of the analysis, such as the summary effect itself or the variance of the summary effect, and considered how these are affected by assumptions of the model. It is important also to step back, and consider the models in a larger context.

The fixed-effect model starts with the assumption that all studies share a common effect size. If we start with that assumption, then the point of the analysis must be to estimate the common effect size.

By contrast, the random-effects model allows that there may be a distribution of true effects. It follows that the first step in the analysis should be to estimate the amount of variation and then use this to inform the direction of the analysis.

If the variation is trivial, then we would focus on reporting the mean and its confidence interval. If the variation is non-trivial, then we might want to address the substantive implications of the variation, but the mean might still be useful as a summary measure. By contrast, if the variation is substantial, then we might want to shift our focus away from the mean and toward the dispersion itself. For example, if a treatment reduces the risk of mortality in some studies while increasing the risk of mortality in others (and the difference does not appear to be due to estimation error), then the focus of the analysis should not be on the mean effect. Rather, it should be on the fact that the treatment effect differs from study to study. Hopefully, it would be possible to identify reasons (differences in the study populations or methods) that might explain the dispersion.

The null hypothesis

The statistical model affects the meaning of the null hypothesis and the test of statistical significance. For the fixed-effect model, the null hypothesis is that the *common* effect size is (for example) zero. For the random-effects model, the null hypothesis is that the *mean* effect size is (for example) zero. While the difference sounds trivial, it actually is critically important—not because of the difference between the common effect size and the mean effect size, but because of the difference in the error terms under the two models.

When we are testing a common effect size there is only one source of error, whereas when we are testing a mean effect size there are two sources of error. If the model calls for one and we use the other, then we are using the wrong value for the standard error. In that case the *Z*-value and *p*-value are, simply put, incorrect.

Statistical power

The discussion above about uncertainty in the effect size estimate under the two models has direct implications for statistical power, the likelihood that the study will yield a statistically significant effect. Statistical power depends on the true effect size, on the criterion for significance, and the precision of the combined effect size, i.e., the standard error [1].

The first two elements, the effect size and the criterion for significance, take on exactly the same form for a fixed-effect or a random-effects analysis. However, the last element, the standard error, differs. For the fixed-effect analysis, as long as the effect size is non-zero, the standard error will eventually approach zero with a large enough sample size (and power will approach 1.0). However, for a random-effects analysis the situation is more complicated. To the extent that the standard error is high because of large within-study variance, we can reduce it (and increase power) by increasing the total sample size (either the sample size within studies or the number of studies). However, to the extent that the standard error is high because of large between-study variance, we can reduce it (and increase power) only by increasing the number of studies [2].

The practical implications are as follows.

First, one should not simply assume that a meta-analysis has good power to detect an effect of practical importance. While power is generally good for a fixed-effect analysis if the total sample size is high, the situation is more complicated for a random-effects analysis. Under the random-effects model, if the between-studies variance is non-trivial and the number of studies is small, it is possible to have tens of thousands of subjects and still have low power.

Second, if we are planning a set of studies prospectively that will be included in a random-effects analysis, we may get much higher power by planning 10 studies with 100 persons each, rather than 5 studies with 200 each. By referring to (19) it should be clear that the within-study error variance will be the same in either case (since the number of subjects is the same), but the between-studies variance will be half as large if the subjects are allocated to 10 studies rather than 5.

Researchers sometimes ask which model yields higher power. This question reflects an assumption that the two models are somehow interchangeable, and that one has the option of selecting the one with higher power. In fact, to ask 'what if we use the fixed-effect model' when the random-effects model is appropriate, is to ask 'what if we ignore part of the error term' when performing the meta-analysis.

Which model should we choose?

Fixed-effect model

It makes sense to use the fixed-effect model if two conditions are met. First, there is good reason to believe that all the studies are functionally identical. Second, our goal is to compute the common effect size, which would not be generalized beyond the (narrowly defined) population included in the analysis. In this paper we presented the example of a sample that was randomly divided into groups. By definition, these groups must share a common mean.

Suppose that a drug company has run five studies to assess the effect of a drug. All studies recruited patients in the same way, used the same researchers, dose, and so on, so all are expected to have the identical effect (as though this were one large study conducted with a series of cohorts). In addition, the regulatory agency only wants to see whether the drug works in this one population. In this example, both conditions are met and the fixed-effect model makes sense.

Random-effects model

In many systematic reviews, the fixed-effect assumption is implausible. When we decide to incorporate a group of studies into a meta-analysis, we assume that the studies have enough in common that it makes sense to synthesize the information, but there is generally no reason to assume that they are 'identical' in the sense that the true effect size is exactly the same in all the studies.

Earlier we presented an example where the goal is to estimate the mean for all colleges in California based on the mean in five colleges selected at random. It is likely that the mean varies from college to college. This is more typical of most real-world syntheses.

Box 1: Estimating τ^2

As explained in the main text, the random-effects computations require an estimate of the between-studies variance, τ^2. We have simply presented an estimate, T^2, and used it in the computations. Here, we outline a process that can be used to estimate τ^2 from the observed data.

Intuitively, it might seem that we could simply compute the variance of the study means and use this value as an estimate of the between-study variance, similar to the approach used in primary studies. This intuition is correct, and if each study had an infinitely large sample size (so that the true effect for each study was known exactly) this approach would work. In fact, though, because the study samples are finite, this approach will not work.

To understand the problem, suppose for a moment that all studies in the analysis shared the same true effect size, so that the (true) between-study variance is zero. Under this assumption, we would not expect the observed effects to be identical to each other. Rather, because of within-study estimation error, we would expect each study mean to fall within *some range* of the common effect.

Now, assume that the true effect size *does* vary from one study to the next. In this case, the observed effects vary from one another for two reasons. One is the real heterogeneity in effect size, and the other is the within-study estimation error. If we want to quantify the real heterogeneity, we need to partition the observed variation into these two components, and then focus on the former.
Conceptually, the approach is a three-step process:

1. We compute the total amount of study-to-study variation actually observed.
2. We estimate how much the observed effects would be expected to vary from each other if the true effect was actually the same in all studies.
3. The excess variation (if any) is assumed to reflect real differences in effect size (that is, the heterogeneity).

One method for estimating τ^2 is the weighted method of moments (or the DerSimonian and Laird method), which proceeds as follows:

$$T^2 = \frac{Q - df}{C}$$

where

$$Q = \sum_{i=1}^{k} W_i (Y_i - M)^2 = \sum_{i=1}^{k} \frac{(Y_i - M)^2}{V_i},$$

$$df = k - 1,$$

k is the number of studies, and

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}.$$

This formula captures the logic outlined above. The statistic Q is a (weighted) sum of squares of the effect size estimates (Y_i) about their mean (M). Q is weighted in such a manner that assigns more weight to larger studies, and this also puts Q on a standardized metric. In this metric, the expected value of Q if all studies share a common effect size is df. Therefore, $Q - df$ represents the excess variation between studies, that is, the part that exceeds what we would expect based on sampling error. Since $Q - df$ is on a standardized scale, we divide by a factor, C, which puts this index back into the same metric that had been used to report the within-study variance, and this value is T^2. If T^2 is less than zero, it is set to zero, since a variance cannot be negative.

The DerSimonian and Laird estimate have the virtue that it is always qualitatively consistent with the heterogeneity test based on the Q statistic (that is, statistically significant heterogeneity is always accompanied by a positive estimate of τ^2). However, it overestimates τ^2 on average and when the number of studies is small, the bias can be substantial. This overestimate of τ^2 can lead to overestimation of the variance of weighted average effect sizes and confidence intervals that are too wide (see [3]). Other methods of estimation of τ^2 are available (such as maximum likelihood estimation or restricted maximum likelihood estimation), but no method can yield very precise estimates of τ^2 when the number of studies is small.

Similarly, suppose that we are working with studies that compare the proportion of patients developing a disease in two groups (say, vaccinated versus placebo). If the treatment works we would expect the effect size (say, the risk ratio) to be similar but not identical across studies. The effect size might be higher (or lower) in studies where the participants are older, or healthier

than others, or when a higher dose is given, and so on. Because studies will differ in the mixes of participants and in the implementations of interventions, among other reasons, there may be different effect sizes underlying different studies.

Or, suppose that we are working with studies that assess the impact of an educational intervention. The magnitude of the impact might vary depending on the other resources available to the children, the class size, the age, and other factors, which are likely to vary from study to study. We might not have assessed these covariates in each study. Indeed, we might not even know which covariates actually are related to the size of the effect. Nevertheless, logic dictates that such factors do exist and will lead to variations in the magnitude of the effect.

In addition, the goal of the analysis is usually to generalize to a range of scenarios. Therefore, if one did make the argument that all the studies used an identical, narrowly defined population, then it would not be possible to extrapolate from this population to others, and the utility of the analysis would be severely limited.

Therefore, in the vast majority of meta-analyses the random-effects model would be the more appropriate choice. It:

- is more likely to fit the actual sampling distribution;
- does not impose a restriction of a common effect size;
- yields the identical results as the fixed-effect model in the absence of heterogeneity;
- allows the conclusions to be generalized to a wider array of situations.

There are four caveats to the above.

The first caveat is that if the number of studies is very small, then the estimate of the between-studies variance (τ^2) will have poor precision. While the random-effects model is still the appropriate model, we lack the information needed to apply it correctly. In this case the reviewer may choose among several options (listed in no particular order), each of them problematic.

Option A is to report the separate effects and not report a summary effect. The hope is that the reader will understand that we cannot draw conclusions about a summary effect size and its confidence interval. The problem is that some readers will revert to other approaches that are sometimes used in narrative reviews (such as counting the number of significant vs non-significant studies) and possibly reach an erroneous conclusion.

Option B is to perform a fixed-effect analysis. This approach would yield a descriptive analysis of the included studies, but would not allow us to make inferences about a wider population. The problem with this approach is that (a) we do want to make inferences about a wider population and (b) readers will make these inferences even if they are not warranted.

Option C is to perform a Bayesian meta-analysis, where the estimate of τ^2 is based on data from outside of the current set of studies. This is probably the best option, but the problem is that relatively few researchers have expertise in Bayesian meta-analysis. Additionally, some researchers have a philosophical objection to this approach.

The second caveat is that, although we define a fixed-effect meta-analysis as assuming that every study has a common true effect size, some have argued that the fixed-effect method is valid without making this assumption [4]. From this perspective, the point estimate of the effect in a fixed-effect meta-analysis is simply a weighted average and does not strictly require the assumption that all studies estimate the same thing. For simplicity and clarity, we adopt a definition of a fixed-effect meta-analysis that does assume homogeneity of effect.

The third caveat is that a random-effects meta-analysis can be misleading if the assumed random distribution for the effect sizes across studies does not hold. A particularly important departure from this occurs when there is a strong relationship between the effect estimates from the various studies and their variances, i.e. when the results of larger studies are systematically different from results of smaller studies. This is the pattern often associated with publication bias, but could in fact be due to several other causes.

The fourth caveat is that, as we move from fixed-effect weights to random-effects weights, large studies lose influence and small studies gain influence. If the effect size is related to sample size, then this will shift the combined result toward the effect size in the smaller studies. Sometimes this is undesirable and it is not clear whether the random-effects result is to be favored over the fixed-effect result where the larger studies will have more of an impact. Many argue that neither analysis is appropriate in this situation and that the source of the relationship should be examined (e.g. [5]).

Mistakes to avoid when selecting a model

In some circles, researchers tend to start with the fixed-effect model, and then switch to the random-effects model if there is a compelling reason to do so. Usually, this means that the test for heterogeneity across studies is statistically significant. This is a bad idea for many reasons.

First, it is a bad idea to use a non-significant heterogeneity test as evidence that the studies share a common effect size. The test for heterogeneity often has poor power, and therefore can be non-significant even if the true level of heterogeneity is substantial.

Second, this approach uses the fixed-effect model as the presumptive model. If we were going to use one model as the default, then the random-effects model is the better candidate because it makes less stringent assumptions about the consistency of effects. Mathematically, the fixed-effect model is really a special case of the random-effects model with the additional constraint that all studies share a common effect size. To impose this constraint is to impose a restriction that is not needed, not proven, and often implausible.

Third, there is no 'cost' to using the random-effects model. Fixed-effect weights are based on within-study variance V, while random-effects weights are based on within-study and between-studies variances, $V+T^2$. It follows that when T^2 is zero the two

models yield the same numerical values for the effect size and its variance. If we use this model and it turns out that the effects are consistent across studies, then results are identical to the fixed-effect model. At the same time, if it turns out that there *is* dispersion among the effects, we can incorporate this into the weighting scheme.

Since the two models yield identical estimates when T^2 is zero, one might say that the fixed-effect model is a special case of the random-effects model. While it is true that the two computational models will yield identical results when T^2 is zero, it may not be helpful to consider the two models in this way. This is because the quantity being estimated in the fixed-effect model is the common effect size in the studies that are observed, whereas the quantity being estimated in the random-effects model is the mean of a hypothetical population of studies, including studies that are not observed.

Fourth, and most fundamentally, is the theme that we have emphasized throughout this paper. The two models really reflect fundamentally different assumptions about the distribution of the data. Rather than start with *either* model as the default, one should select the model based on their understanding about whether or not the studies share a common effect size and on their goals in performing the analysis.

Weighting schemes

In this paper we have concentrated on the use of inverse-variance weights for two reasons. First, these are commonly used. Second, inverse-variance weights are especially well suited to highlighting the differences between fixed-effect and random-effects models because these differences are apparent in the definition of the variance. However, inverse-variance weights are not the only weighting scheme available. Other schemes include Mantel–Haenszel weights for binary data and weighting by sample size for correlations [6]. The differences among these approaches (including the inverse-variance approach) are relatively minor. They all assume that all studies in the analysis either share a common effect size (fixed-effect model) or are a random sample from a universe of potential studies (random-effects model). If either assumption is correct, then *any* weighting scheme will be unbiased. The schemes listed above have the advantage of being efficient (yielding a precise estimate) because they assign more weight to the more precise studies.

Some recent high-profile papers have argued for a fundamentally different approach to meta-analysis that assigns equal weight to all studies. Bonett [7, 8] argues against the fixed-effect model because it appears implausible and against the random-effects model because the studies in the analysis are not really a random sample of all possible studies. He also has concerns over the poor estimation of τ^2 when there is a small number of studies. Bonett argues that it would be better to report the unweighted mean for the studies at hand. Shuster [9] also argues for the unweighted mean, but primarily due to bias in the results of all weighted methods when there is correlation between the estimates and the weights.

These suggestions to use equal weighting constitute a radical departure from standard meta-analytic practice and will likely be the topic of some debate. Many may feel intuitively uncomfortable about ignoring the sample sizes or precisions of the different studies when computing a combined effect. An unweighted approach also carries some potential technical shortcomings. In particular, the simple mean may be a very inefficient estimator (the summary effect will have a large variance), particularly in cases where the variance differs substantially from study to study. In such cases, a weighted estimate may *still* be preferable to an unweighted estimate—while the weighted estimate may be biased, it might still be more accurate in the sense that it will have smaller mean-squared error than the unweighted estimate.

Moderator variables

While the random-effects model takes account of variation in effect sizes, it simply incorporates this variation into the weighting scheme—it makes no attempt to explain this variation. This is appropriate when the variation is assumed to be random, in the sense that it depends on unknown factors. By contrast, if we anticipate that some of the variance can be explained by specific variables, then we can study the relationship between these putative moderators and effect size (e.g. [10]). This has the effect of reducing the unexplained (residual) between-studies variation.

For example, suppose the analysis includes 20 studies, where 10 employed one variant of the intervention while 10 employed another variant. We can perform a subgroup analysis, computing a summary effect within each set of 10 studies, and then comparing the two summary effects. This allows us to partition the total variance (of all effects vs the overall mean) into variance between the two subgroup means and variance within the subgroups. In keeping with the overall theme of this paper, we can assume either a fixed-effect or a random-effects model within each of the subgroups, depending on whether or not we assume that all true effects within a subgroup are identical [11–13].

Similarly, if we expect that the variation in effect sizes is related to a continuous variable, we can use meta-regression to study this relationship. Again, we can apply either a fixed-effect or a random-effects variant of this approach, depending on whether or not we assume that all studies with the same covariate values will have the same true effect [11, 12].

Issues not covered in this paper

In keeping with the goals of this paper we have focused exclusively on the nature, differences and implications of fixed-effect vs random-effects models. We have not addressed many other issues related to estimating, analyzing and interpreting effect sizes. Two of those issues that are more closely related to the topic of this paper are outlined below. Our intent here is to alert the reader to these issues and provide some references.

In meta-analysis, as in primary studies, the analysis can focus on either a test of the null hypothesis or a report of the effect size with confidence intervals. While these are complementary to each other (the *p*-value will generally fall under 0.05 if and only

if the 95% confidence interval does not include the null value), the two approaches address very different questions. The former addresses the question, 'Can we reject the null hypothesis?' The latter addresses the questions, 'What is the size of the effect' and 'How precise is our estimate?' There are cases where the goal of the analysis *really is* to test the null and in those cases the analysis should focus on the test of significance. In the majority of cases, however, the viability of the null hypothesis is of little, if any, interest and the goal is to estimate the magnitude of the effect. In those cases, the analysis should focus on the estimate of the effect size and the precision of that estimate [14–16].

In this paper we have taken the effect estimates from the individual studies at face value. That is, we have assumed that the effect estimates are unbiased. In reality, effect sizes may be biased due to flaws in the design or conduct of the studies from which they are derived. Furthermore, measurement error in independent or dependent variables can reduce the magnitude of the effect sizes compared with the magnitude that would have been observed if the variables had been measured without error. For example, measurement error attenuates correlation coefficients, although they can be adjusted (disattenuated) to correct for those effects [17]. Similarly, measurement error in the dependent (outcome) variable attenuates standardized mean differences, and these can be adjusted to correct for that effect [18]. In some cases the effects of measurement error can be substantial and, therefore, researchers may prefer to work with effect size estimates corrected for the effects of measurement error [6, 19]. The methods we have described can be applied to effect size estimates and variances that have been corrected for measurement error or other biases. However, an additional complexity arises from the fact that researchers who work with such corrected effect size estimates often also use methods other than the inverse-variance approach for combining the estimates in the meta-analysis. A comparison of the Schmidt–Hunter approach and the inverse-variance approach is provided in Schulze [20] and Borenstein *et al.* [12].

Summary

- The fixed-effect model is based on the assumption that all studies in the meta-analysis share a common (true) effect size. By contrast, the random-effects model allows that the true effect size may vary from study to study.
- Under the fixed effect model, the summary effect is an estimate of the effect which is common to all studies in the analysis. Under the random-effects model the summary effect is an estimate of the mean of a distribution of true effects.
- Study weights are more uniform (similar to one another) under the random-effects model than under the fixed-effect model. Large studies are assigned less relative weight and small studies are assigned more relative weight as compared with the fixed-effect model.
- The standard error of the summary effect and, therefore, the confidence intervals for the summary effect are wider under the random-effects model than under the fixed-effect model.
- Under the fixed effect model the only source of variation is the within-study estimation error. With a large enough sample size, accumulated across studies, this source of variation will diminish and the common effect size will be estimated precisely.
- Under the random-effects model there are two sources of variation, within-study estimation error variance and between-studies variance. With a large enough sample size, accumulated across studies, the effect of first source of variation will diminish. However, if the between-studies variance is substantial, the only way to obtain good precision is to increase the number of studies. If we increase the sample size in a few studies we may know the effect size in those studies very precisely, but still not have a precise estimate of the mean across all studies.
- The selection of a model should be based solely on the question of which model fits the distribution of effect sizes and thus takes account of the relevant source(s) of error. When studies are gathered from the published literature, the random-effects model is generally a more plausible match.
- The strategy of starting with a fixed-effect model and then moving to a random-effects model if the test for heterogeneity is significant relies on a flawed logic and should be strongly discouraged.
- While the random-effects model is often the appropriate model, there are cases where it cannot be implemented properly because there are too few studies to obtain an accurate estimate of the between-studies variance.

Appendix A

This appendix includes the computations for the motivational example.

The computations for the fixed-effect model are shown in Table AI.

Using the numbers from this table and formula (6) the summary effect size is

$$M = \frac{402.350}{4.000} = 100.588.$$

Then using (9) and (10), the variance and standard error are

$$V_M = \frac{1}{4.000} = 0.250$$

and

$$\mathrm{SE}_M = \sqrt{0.250} = 0.500.$$

Table AI. Computations for the fixed-effect model.

	M	SD	*N*	*V*	*W*	*WY*	WY^2	W^2
Study A	99.100	20.000	200	2.000	0.500	49.550	4910.405	0.250
Study B	101.200	20.000	200	2.000	0.500	50.600	5120.720	0.250
Study C	101.800	20.000	800	0.500	2.000	203.600	20726.480	4.000
Study D	98.100	20.000	200	2.000	0.500	49.050	4811.805	0.250
Study E	99.100	20.000	200	2.000	0.500	49.550	4910.405	0.250
Total					4.000	402.350	40479.815	5.000

Table AII. Computations for the random-effects model.

	M	SD	*N*	V_{within}	T^2	V_{total}	*W*	*WY*	WY^2	W^2
Study A	99.100	20.000	200	2.000	1.6125	3.612	0.277	27.433	2718.563	0.077
Study B	101.200	20.000	200	2.000	1.6125	3.612	0.277	28.014	2835.001	0.077
Study C	101.800	20.000	800	0.500	1.6125	2.112	0.473	48.189	4905.676	0.224
Study D	98.100	20.000	200	2.000	1.6125	3.612	0.277	27.156	2663.975	0.077
Study E	99.100	20.000	200	2.000	1.6125	3.612	0.277	27.433	2718.563	0.077
Total							1.581	158.224	15841.778	0.531

The 95% confidence interval, per (11) and (12), is

$$\mathrm{LL}_M = 100.588 - 1.96 \times 0.500 = 99.608$$

to

$$\mathrm{UL}_M = 100.588 + 1.96 \times 0.500 = 101.568.$$

To test the null hypothesis that the common true score is 100, we would use (13) to compute

$$Z = \frac{100.588 - 100.000}{0.500} = 1.175.$$

For a one-tailed test the corresponding *p*-value is given in Microsoft Excel by =1−NORMSDIST(ABS(Z)), which equals 0.120. For a two-tailed test the corresponding *p*-value is given in Excel by =2*(1-NORMSDIST(ABS(Z))), which equals 0.240.

The computations for the random-effects model are shown in Table AII.

First, we compute T^2 using the formulas in Box 1, as follows:

$$T^2 = \frac{8.434 - 4}{2.750} = 1.6125$$

where

$$Q = 40479.815 - \frac{402.350^2}{4.000} = 8.434,$$

$$df = 5 - 1 = 4,$$

and

$$C = 4.000 - \frac{5.000}{4.000} = 2.750.$$

Using the numbers from this table and formula (6) the summary effect size is

$$M = \frac{158.224}{1.581} = 100.101.$$

Then using (9) and (10), the variance and standard error are

$$V_M = \frac{1}{1.581} = 0.633$$

and

$$\mathrm{SE}_M = \sqrt{0.633} = 0.795.$$

The 95% confidence interval, per (11) and (12), is

$$LL_M = 100.101 - 1.96 \times 0.795 = 98.542$$

to

$$UL_M = 100.101 + 1.96 \times 0.795 = 101.660.$$

To test the null hypothesis that the common true score is 100 we would use (13) to compute

$$Z = \frac{100.101 - 100.000}{0.795} = 0.127.$$

For a one-tailed test the corresponding *p*-value is given in Microsoft Excel by =1−NORMSDIST(ABS(Z)), which equals 0.449. For a two-tailed test the corresponding *p*-value is given in Excel by =2*(1-NORMSDIST(ABS(Z))), which equals 0.899.

These numbers provide a concrete example of the principles outlined in the paper. The weight assigned to each study, and the sum of weights across studies, is smaller under the random-effects model. This leads to a larger variance and a wider confidence interval.

The study weights are more similar under the random-effects model. For example, under the fixed-effect model the relative weights assigned to the smaller studies (A, for example) and the larger study (C) are 12.5 vs 50%. By contrast, under the random-effects model these relative weights are 17.5 vs 30%. In this example the large study (C) happens to have a relatively high mean. Therefore, the combined mean is higher under the fixed-effect model (where this study has more influence) and lower under the random-effects model (where this study has a more modest influence).

Acknowledgements

This paper is adapted, with some additions, from chapters in Introduction to Meta-Analysis (Borenstein, Hedges, Higgins and Rothstein, 2009). The authors gratefully acknowledge the financial support provided by the following grants. From the National Institutes of Health (to MB): Combining data types in meta-analysis (AG021360), Publication bias in meta-analysis (AG20052), Software for meta-regression (AG024771), and Forest plots for meta-analysis (DA019280). From the IES (to LH): Representing and combining the results of randomized trials in education (R305U080002). JH is supported by MRC grant U.1052.00.011. HR is supported by a fellowship leave from Baruch College, CUNY.

References

1. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates: Hillside, NJ, 1987.
2. Hedges LV, Pigott TD. The power of statistical tests in meta-analysis. *Psychological Methods* 2001; **6**:203–217.
3. Hedges LV, Vevea JL. Fixed and random effects models in meta-analysis. *Psychological Methods* 1998; **3**:486–504.
4. Early Breast Cancer Trialists' Collaborative Group. *Treatment of Early Breast Cancer; Volume 1*: *Worldwide Evidence 1985–1990*. Oxford University Press: Oxford, U.K., 1990.
5. Sterne JAC, Egger M, Moher D (eds). Addressing reporting biases. In *Cochrane Handbook for Systematic Reviews of Intervention*, Higgins JPT, Green S (eds), Chapter 10. Wiley: Chichester, U.K., 2008.
6. Hunter JE, Schmidt FL. *Methods of Meta-analysis*: *Correcting Error and Bias in Research Findings*. Sage: Thousand Oaks, CA, 2004.
7. Bonett DG. Meta-analytic interval estimation for bivariate correlations. *Psychological Methods* 2008; **13**:173–189.
8. Bonett DG. Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods* 2009; **14**:225–238.
9. Shuster JJ. Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine* 2010; **29**:1259–1265.
10. Hedges LV, Olkin I. *Statistical Methods for Meta-analysis*. Academic Press: San Diego, CA, 1985.
11. Lipsey M, Wilson D. *Practical Meta-analysis*. Sage: Thousand Oaks, CA, 2001.
12. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-analysis*. Wiley: Chichester, 2009.
13. Cooper H, Hedges LV, Valentine JC. *The Handbook of Research Synthesis* (2nd edn). Russell Sage Foundation: New York, 2009.
14. Borenstein M. The case for confidence intervals in controlled clinical trials. *Controlled Clinical Trials* 1994; **15**:411–428.
15. Borenstein M. The shift from significance testing to effect size estimation. *Comprehensive Clinical Psychology*, Bellack AS, Herson M (eds). Pergamon: New York, 2000; 313–349.
16. Ziliak ST, McCloskey DN. *The Cult of Statistical Significance*: *How the Standard Error Costs us Jobs, Justice, and Lives*. University of Michigan Press: Ann Arbor, 2008.
17. Spearman C. The proof and measurement of the association between two things. *American Journal of Psychology* 1904; **15**:72–101.
18. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 1981; **6**:107–128.
19. Schmidt FL, Hunter JE. Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology* 1977; **62**:529–540.
20. Schulze R. *Meta-analysis*: *A Comparison of Approaches*. Hogrefe and Huber Publishers: Cambridge, MA, 2004.