Journal of Educational and Behavioral Statistics
April 2010, Vol. 35, No. 2, pp. 215–247
DOI: 10.3102/1076998609346961

How Many Studies Do You Need? A Primer on Statistical Power for Meta-Analysis

Jeffrey C. Valentine
University of Louisville

Therese D. Pigott
Loyola University-Chicago

Hannah R. Rothstein
Baruch College

In this article, the authors outline methods for using fixed and random effects power analysis in the context of meta-analysis. Like statistical power analysis for primary studies, power analysis for meta-analysis can be done either prospectively or retrospectively and requires assumptions about parameters that are unknown. The authors provide some suggestions for thinking about these parameters, in particular for the random effects variance component. The authors also show how the typically uninformative retrospective power analysis can be made more informative. The authors then discuss the value of confidence intervals, show how they could be used in addition to or instead of retrospective power analysis, and also demonstrate that confidence intervals can convey information more effectively in some situations than power analyses alone. Finally, the authors take up the question "How many studies do you need to do a meta-analysis?" and show that, given the need for a conclusion, the answer is "two studies," because all other synthesis techniques are less transparent and/or are less likely to be valid. For systematic reviewers who choose not to conduct a quantitative synthesis, the authors provide suggestions for both highlighting the current limitations in the research base and for displaying the characteristics and results of studies that were found to meet inclusion criteria.

Keywords: *meta-analysis; research methodology; statistics*

Scholars undertaking a literature review can have a variety of goals. Some reviews, for example, are meant to provide insight into important theories or themes in a literature. Other reviews have more specific aims. For example,

Please address correspondence to Jeff Valentine, 309 CEHD, Louisville, KY 40290; e-mail: jeff.valentine@louisville.edu

many literature reviews are undertaken with the goal of describing a relationship, such as whether some intervention is effective at bringing about some change or whether two variables are associated. In these cases, a systematic review is often the preferred method for locating studies and analyzing their results. A systematic review can be defined as a systematic and transparent approach to the collection and evaluation of a literature on a specific research question. Using this definition, one can draw a distinction between the terms "systematic review" and "meta-analysis," which are often used interchangeably. We define meta-analysis as the quantitative analysis of the results of multiple studies, and note that a systematic review need not include a meta-analysis nor must a meta-analysis be based on studies located through a systematic reviewing process (in most cases, though, a meta-analysis should be based on a systematic review).

Potential reviewers are often faced with decisions about whether literature has enough information (i.e., studies) to justify the effort associated with conducting a systematic review. In addition, once a systematic review has been undertaken, questions may arise concerning the feasibility and appropriateness of synthesizing the data quantitatively (i.e., conducting a meta-analysis). Finally, if a meta-analysis is conducted, researchers are faced with the issue of how to appropriately interpret the statistical result. This is especially problematic if a test of statistical significance reveals a statistically nonsignificant relationship. The aim of this article is to review, illustrate, and critically examine three of the principal approaches to the decision about when to include a meta-analysis in a systematic review. Two of these, power analysis and an examination of confidence interval width, are primarily statistical (although both require the researcher to make judgments and estimates). A third approach, the assessment of the heterogeneity of a body of studies, is somewhat more subjective but can be given a statistical foundation. For each approach, we will review the factors that need to be taken into account and provide a running illustrative example, using both fixed and random effects models. The final section of the article will address the concern that "the data are too heterogeneous for meta-analysis." Although we believe that this is less often the case than sometimes claimed, we offer some options for presenting study results when reviewers decide that they cannot quantitatively synthesize the literature. Our ultimate goal for this article is to help further the evolution of literature reviewing, so that education researchers and policymakers can have the best possible information as they design, fund, and implement educational interventions.

Prospective Statistical Power Analysis

A commonly given reason for conducting a meta-analysis is that a collection of related studies have higher statistical power than any single one of those studies (Cohn & Becker, 2003). Even so, it is still reasonable to consider the statistical power of the collection of studies, because it may not be high enough to

have a reasonable probability of detecting the effect size of interest assuming it actually exists. As such, scholars thinking about conducting a systematic review (and perhaps, their funders or dissertation advisors) may be interested in investigating the likelihood that the research question is supported by a sufficient body of research to support a quantitative analysis of study results; in other words, they may wish to estimate the statistical power of the planned meta-analysis. Details for computing the power for meta-analytic tests are given in Hedges and Pigott (2001, 2004). These authors show that statistical power in meta-analysis is conceptually similar to statistical power in primary studies. In both cases statistical power is a function of the estimated population effect size, the Type I error rate, and sample size. Of course, the relations among these variables are such that increasing any one of them (while holding the others constant) serves to increase power.

Reviewers also need to choose to analyze their data using fixed effects or random effects assumptions. Reviewers use fixed effects assumptions when they believe that the results of studies vary only due to random sampling error and identifiable covariates; in other words, they believe that studies are all estimating a single population effect size. Random effects assumptions are used when reviewers believe that study results are affected by random sampling error, identifiable covariates, and unidentifiable sources of variability (i.e., unmeasured covariates); in other words, they believe that studies come from a distribution of population effect sizes. Fixed effects analyses allow for inferences only to studies that are highly similar to the ones included in the meta-analysis, while random effects analyses allow for inferences to a wider range of studies (Hedges & Vevea, 1998).

Furthermore, as with prospective power analysis in a primary study, prospective power analysis in a meta-analysis depends on the reviewers' estimates regarding certain parameters that are unknown before data collection begins. Specifically, for a fixed effects prospective power analysis reviewers need to estimate (a) the smallest effect size that is meaningful in the context of the research question (or alternatively, they could estimate the effect size that they would expect to find) and (b) the variance of this overall average effect size. This latter quantity may be approximated in a few different ways. If the reviewers have an estimate about the likely confidence interval of the mean effect size, that could be used to estimate the variance. This is unlikely to be the case, however. A second (and more practical) strategy involves reviewer conjectures regarding (a) the number of studies that will meet inclusion criteria (k) and (b) the average within-study sample size (n). These quantities can be roughly estimated based on the researchers' general familiarity with the research literature he or she is reviewing. Alternatively, the researchers could carry out a scoping review (Torgerson, 2003), which is based on a limited literature search and is designed to provide a preliminary assessment of the size and nature of the research literature relevant to a specific topic. Scoping reviews are increasingly recommended as a means of determining the feasibility and relevance of a full-scale systematic review (Arksey & O'Malley, 2005).

Reviewers also need to specify the Type I error rate threshold they will use and whether they will conduct a one-tailed (directional) statistical test or a two-tailed (nondirectional) test. For good or ill, in the educational and social sciences the desired Type I error rate is typically set at $\alpha = .05$. Directional tests are usually chosen when the researchers have a strong a priori hypothesis about the nature of the relationship. Alternatively, Cohen (1988) argues that one-tailed tests should be used when a statistically significant finding in the opposite direction to the one hypothesized would have the same implications as a null finding. For example, a new reading intervention might be more effective, no more effective than the standard strategies, or harmful relative to the standard strategies. If it is either no more effective or no more harmful, we would not recommend the new intervention, and according to Cohen, a one-tailed test would be justified. When, however, a statistically significant finding in the opposite direction has a different meaning than a null result, a two-tailed test should be used, and the power calculations must take this into account. An example might be when a researcher is comparing the relative effectiveness of two well-established reading interventions.

Conducting a random effects power analysis for a test that a relationship is not zero requires an additional conjecture: the size of the between-studies variance component, τ^2. Strategies for estimating value will be presented later. Although we believe that the random effects model is appropriate in most cases in education research, we will first focus on the fixed effect case for the sake of pedagogical simplicity and ease of illustration. The examples below are based on the standardized mean difference effect size, for which the population parameter is denoted as δ and the sample statistic is denoted as d, which is defined as

$$d = \frac{M_T - M_C}{s_p}, \tag{1}$$

where M_T = the mean of the treatment group, M_C = the mean of the comparison group, and s_p = the pooled standard deviation (i.e., the square root of the weighted mean of the variances of the two groups). Hedges and Pigott (2001, 2004) provide examples using other effect size metrics. For all the common metrics, the power computations are identical.

Prospective statistical power for the fixed effects mean effect. As an example, assume a review team is interested in the effectiveness of programs that aim to prevent depression among adolescents. For self-report measures of depressive symptoms, they have a sense that a standardized mean difference effect size of $\delta = 0.15$ would be necessary for any effect to be clinically meaningful. Furthermore, their preliminary searches suggest that the studies in this area tend to be rather small (e.g., 20 participants per condition), but they expect that a large number of studies will meet their inclusion criteria (e.g., 40). Following convention, they plan to use a Type I error rate of $\alpha = .05$, but, because a statistically

significant effect in the opposite direction (i.e., a finding that the program harms participants relative to the comparison condition) would be treated like a null finding (i.e., the new program would not be adopted), they plan on conducting a directional test.

In meta-analysis, the procedures for assessing power for determining whether the overall effect is different from zero are based on the Z statistic. To compute power, we need to make assumptions about the "typical" study in a meta-analysis. For example, when testing whether an overall effect size is different from zero, we use the test given by

$$Z = (\overline{\mathrm{ES}} - 0)/\sqrt{v_\bullet}. \tag{2}$$

$\overline{\mathrm{ES}}$ is the weighted mean effect size in this example and $v_\bullet$ is the value of the variance of the weighted mean effect size, given by

$$v_\bullet = \frac{1}{\sum_{i=1}^{k} (1/v_i)}, \tag{3}$$

where k is the anticipated number of studies. Here, v_i is the variance for the effect size from the ith study (defined below in Equation 4); the inverse of this value is the weight accorded to a study in a meta-analysis. When planning a meta-analysis, we do not know the value of the variance of the mean effect size because we need the variance from each study's effect size. Instead, we can suggest characteristics of a "typical" study we may find in a synthesis such as within-study sample sizes. Combined with a substantively important value for the effect size, $\overline{\mathrm{ES}}$, we can pose a "typical" value for the weighted mean effect size variance. If our "typical" study compares the outcomes of an intervention and a control group (and therefore intervention effects are most naturally described in terms of the standardized mean difference), then the value for our "typical" study's effect size variance is given by v:

$$v = \frac{n_\mathrm{T} + n_\mathrm{C}}{n_\mathrm{T} n_\mathrm{C}} + \frac{\overline{\mathrm{ES}}^2}{2(n_\mathrm{T} + n_\mathrm{C})} = 1/w, \tag{4}$$

where n_T = the hypothesized size of the typical treatment group, n_C = the hypothesized size of the typical comparison group, $\overline{\mathrm{ES}}$ = the effect size that the reviewers would like to be able to detect, and w is the weight the "typical" study receives in a meta-analysis. If we assume that all studies have similar sample sizes, then we have a value of $v_\bullet$, the variance of the effect size, equal to

$$v_\bullet = \frac{1}{\sum_{i=1}^{k} (1/v)} = \frac{1}{k/v} = \frac{v}{k}. \tag{5}$$

When the overall effect is statistically significantly different from zero, the Z statistic has a normal distribution with a mean equal to

$$\lambda = (\overline{ES} - 0)/\sqrt{v_\bullet} = (\overline{ES} - 0)\Big/\sqrt{v/k} \tag{6}$$

and a variance of 1. When we reject the null hypothesis, our Z statistic is no longer distributed as a standard normal with mean 0 and variance 1. Instead, the mean is given by λ, and the variance is 1. This value (λ) is known as the noncentrality parameter and expresses the extent to which the null hypothesis is false. To calculate the prospective power of a weighted mean effect size, $\overline{ES}$, with variance equal to $v_\bullet$, the one-sided power of the test that the mean effect is greater than zero is given by

$$p = 1 - \Phi(c_\alpha - \lambda), \tag{7}$$

where c_α is the $100(1 - \alpha)$ critical value for the standard normal distribution and $\Phi(x)$ is the standard normal cumulative distribution function. For $\alpha = .05$ and a one-tailed test, $c_\alpha = 1.64$ (for a two-tailed test, $c_\alpha = 1.96$ at $\alpha = .05$). The standard normal cumulative distribution function is given in several spreadsheets and statistical packages. For example, in Microsoft's Excel program, the command "=normsdist(x)" (typed in without the quotation marks) will return the standard normal cumulative distribution function for the value of x. It is also given in statistical textbooks that present a z table (e.g., Howell, 1992).

To estimate the variance of the overall average effect size, the researchers can use their estimates about the number of studies that will meet inclusion criteria (i.e., $k = 40$) and their estimate of the typical within-study sample size (i.e., $n = 20$ per condition). They would then use Equation 4 to compute the typical study's effect size variance:

$$v = \frac{20 + 20}{(20)(20)} + \frac{.15^2}{2(20 + 20)} = \frac{40}{400} + \frac{.0025}{80} = .1003.$$

The reviewers could then use Equation 5 to compute the variance of the effect size,

$$v_\bullet = \frac{.1003}{40} = .0025$$

and then Equation 6 to compute the value of λ:

$$\lambda = \frac{(.15 - 0)}{\sqrt{.0025}} = 3.$$

Finally, the reviewers would use Equation 7 to compute fixed effects prospective power using these assumptions, which yields:

$$p = 1 - \Phi(1.64 - 3) = 1 - \Phi(-1.36) = 1 - .087 = .913.$$

TABLE 1
Illustration of Fixed Effects Prospective Power (One-Tailed) as a Function of Different Assumptions About Review Parameters

Within-Study Sample Size (Per Group)	Number of Studies to Be Included	Effect Size to Detect	Power
10	40	0.15	0.68
20	40	0.15	0.91
30	40	0.15	0.98
40	40	0.15	≈1.0
20	10	0.15	0.44
20	25	0.15	0.77
20	40	0.15	0.91
20	65	0.15	0.98
20	40	0.05	0.26
20	40	0.15	0.91
20	40	0.25	≈1.0
20	40	0.35	≈1.0

Note: Effect size is expressed in standardized mean difference units.

In other words, for this example, the estimated statistical power for the test that the overall effect size is not zero is approximately .91. This result suggests that a meta-analytic test that the mean effect size is greater than zero will correctly reject the false null hypotheses about 91% of the time, given the assumptions specified for this example. Table 1 presents a demonstration of how prospective fixed effects power varies as a function of differing assumptions about the parameters of the meta-analysis.

Prospective power for the random effects mean effect. Computing the random effects power for the test of the null hypothesis that the treatment effect is zero is complicated somewhat by the need for an estimate of the between-studies variance component, which reflects the additional uncertainty that arises when included studies are no longer thought of as estimating the same underlying population parameter. The statistic τ is an estimate of the standard deviation of the population effect sizes, and τ^2 is its variance. As such, τ is expressed in the same metric as the effect size, and τ^2 is expressed in the same (squared) effect size metric. Like all standard deviations and variances, both statistics can take on values ranging from zero to infinity.

In many cases, even if the reviewer is quite familiar with the literature, he or she will not have a good means of estimating the between-studies variance component. Unfortunately, an intuitively reasonable strategy, namely consulting other meta-analyses in the same general field, is unlikely to be helpful because these values are not routinely reported in meta-analyses, thus depriving reviewers

of a normative source of data on the likely range of values of τ^2 for the given outcome in the research context. When previous experience or prior literature reviews do not provide a rough guide regarding the likely range of τ^2, the reviewers might consider as a starting point the breadth of their research question, which gives an indication of the diversity of studies that they intend to include. For example, some researchers narrowly tailor the parameters of their synthesis. These researchers might adopt relatively stringent criteria that limit the variability of the independent variable, the dependent variables, and/or the methodology used in the included studies. As an example, a researcher interested in the effects of after-school prevention programming might limit studies to those that (a) target students at risk of developing depression, (b) are delivered by trained clinicians, and (c) use random assignment of students to conditions. In these cases, a relatively small value of τ^2 could be expected. Alternatively, some researchers adopt relatively inclusive criteria that serve to increase the variability among the included studies. Again, in the example of a researcher interested in the effects of school-based programs aimed at preventing adolescent depression, the researcher might include studies that (a) target at-risk students or are delivered to students regardless of risk, (b) use clinicians, paraprofessionals, and/or teachers to deliver the program, and (c) use either random assignment or some other (nonrandom) mechanism to place students into groups. In these cases, relatively larger values of τ^2 might be expected due to the diversity of methods used in the component studies.

Most researchers who are used to working with power analysis have come to rely on conventions for determining the values of small, medium, and large effects (a practice we discourage for helping determine the *meaningfulness* of an effect; see Cooper, 2008). Regardless, there are no analogous general conventions for small, medium, and large values of τ^2; at least not in the social sciences or in education research. As such, this determination is probably best guided by familiarity with a particular substantive area. When past experience in the area is not sufficient to guide the decision, we suggest that prospective meta-analysts use the work of Higgins and colleagues (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003) to guide their selection of values of τ^2. These authors developed an index, known as I^2, to describe the degree of heterogeneity in the outcomes of a set of studies. I^2 is defined as

$$I^2 = \frac{Q - (k - 1)}{Q} \times 100, \tag{8}$$

with Q (see Hedges & Olkin, 1985) defined as

$$Q = \sum w_i(\mathrm{ES}_i - \overline{\mathrm{ES}})^2 \tag{9}$$

and all other terms defined as above. I^2 is related to the Birge ratio (1932), which is the ratio of between-studies variation to within-study sampling error. I^2 can

FIGURE 1. *The pie charts represent I^2 values of 25%, 50%, and 75%, reading top to bottom. Light shaded slices represent between-study variability and dark shaded slices represent within-study variability. In the top figure, the ratio of between-study to within-study variability is 1:3, which implies that the between-studies variance component is $\tau^2 = 1/3(\nu_\bullet)$ or $\tau^2 = .33(\nu_\bullet)$. See Equation 10.*

take on negative values. By convention, these are set to 0. Thus, defined in this manner, I^2 can take on values ranging from 0% to 100%. Based on an analysis of I^2 values from more than 500 meta-analyses in medicine, Higgins et al. (2003) suggest that an I^2 value of 25% represents a small degree of heterogeneity, while 50% represents a moderate degree, and 75% represents a large degree of heterogeneity. One way to think about I^2 is that it represents the proportion of total variation in effect sizes that is attributable to between-study variance, τ^2. Thus, Equation 8 can be rewritten as

$$I^2 = \frac{\tau^2}{\tau^2 + \nu_\bullet}. \tag{10}$$

As can be seen in the first pie chart in Figure 1, if a small degree of heterogeneity is defined as meaning that 25% of the total variability in effect sizes is attributable to between-study variation, it suggests that there exists one third as much between-study variability as there is within-study variability, or $\tau^2 = .33(\nu_\bullet)$. Continuing with Higgins and colleagues' recommendations, $\tau^2 = 1.0(\nu_\bullet)$ is a moderate degree of heterogeneity, and $\tau^2 = 3.0(\nu_\bullet)$ is a large degree of heterogeneity. These are represented in the second and third pie charts in Figure 1, respectively. Therefore, in conjunction with the common variance estimated in the fixed effects prospective power analysis, random effects power can be estimated given the assumption of large, moderate, or small degrees of heterogeneity.

Following Hedges and Pigott (2001), we will notate random effects estimates with an asterisk. Therefore, to test whether the overall random effects estimate is statistically significantly different from zero, we rewrite Equation 6 to

$$\lambda^* = \frac{\overline{\mathrm{ES}} - 0}{\sqrt{\nu_\bullet^*}}, \tag{11}$$

with $\nu_\bullet^*$, the random effects variance associated with the weighted mean effect size, defined as

$$\nu_\bullet^* = \frac{1}{\sum_{i=1}^{k} (1/\nu^*)} = \frac{1}{k/\nu^*} = \frac{\nu^*}{k} \tag{12}$$

and the "typical" sampling variance of the random effects estimate of an effect size is defined as

$$v^* = \left(\frac{n_T + n_C}{n_T n_C} + \frac{\overline{ES}^2}{2(n_T + n_C)}\right) + \tau^2 = v + \tau^2. \quad (13)$$

The one-sided power for the random effects tests that the mean effect size is greater than zero is given by

$$p = 1 - \Phi(c_\alpha - \lambda^*). \quad (14)$$

To estimate the random effects variance of the overall average effect size, the researchers can use their estimates about the number of studies that will meet inclusion criteria (i.e., $k = 40$) and their estimate of the typical within-study sample size (i.e., $n = 20$ per condition), and their estimate of the between-studies variance component τ^2. Assume that the researchers estimated that the studies would exhibit a moderate degree of heterogeneity. From our earlier example, we computed the value of $v = .1003$. Assuming a moderate degree of heterogeneity, estimated $\tau^2 = 1.0(v) = 1.0(.1003) = .1003$. Therefore, the reviewers would use Equation 13 to compute the value of the typical sampling variance of the random effects estimate of the effect size, which is $v^* = 0.1003 + 0.1003 = 0.2006$. From Equation 12, the variance for the weighted mean effect size using random effects is $v^*_\bullet = 0.2006/40 = .005$. The reviewer would then use Equation 11 to compute the value of λ, which here is

$$\lambda^* = (\overline{ES} - 0)/\sqrt{v^*_\bullet} = .15/\sqrt{.005} = 2.12.$$

Finally, to compute random effects power using Equation 14

$$p = 1 - \Phi(c_\alpha - \lambda^*) = 1 - \Phi(1.64 - 2.12) = 1 - \Phi(-.48) = 1 - .32 = .68.$$

This example suggests that in this case, the power to detect a true population effect of $\delta = 0.15$ is moderate to high given a random effects model and assuming a moderate degree of heterogeneity. Note that the random effects power is usually lower than the fixed effects power given an estimated $\tau^2 > 0$ (in some unusual situations, random effects power can be greater than fixed effects power). Estimated prospective power for a random effects analysis would be higher if a smaller degree of heterogeneity were expected and lower if a greater degree of heterogeneity were expected. Table 2 demonstrates how prospective random effects power varies as a function of differing assumptions about the parameters of the meta-analysis, and Figure 2 provides a visual comparison of the statistical power of a meta-analysis under different error models and assumptions (fixed effects, random effects assuming a small degree of heterogeneity, and random effects assuming a large degree of heterogeneity).

TABLE 2
Illustration of Random Effects Statistical Power (One-Tailed) as a Function of Different Assumptions About Review Parameters

Within-Study Sample Size (per group)	Number of Studies to Be Included	Effect Size to Detect	Degree of Heterogeneity	Power
10	40	0.15	Moderate	0.54
20	40	0.15	Moderate	0.68
30	40	0.15	Moderate	0.75
40	40	0.15	Moderate	0.79
20	10	0.15	Moderate	0.28
20	25	0.15	Moderate	0.51
20	40	0.15	Moderate	0.68
20	65	0.15	Moderate	0.83
20	40	0.05	Moderate	0.18
20	40	0.15	Moderate	0.68
20	40	0.25	Moderate	0.97
20	40	0.35	Moderate	≈1.0
20	40	0.15	Large	0.44
20	40	0.15	Moderate	0.68
20	40	0.15	Small	0.83

Note: Effect size is expressed in standardized mean difference units. Following the example in the text (see "Prospective power for the random effects mean effect" above), the values of τ^2 used to represent large, moderate, and small degrees of heterogeneity are 3.0, 1.0, and 0.33, respectively.

Power of the test of homogeneity. The finding that a mean effect size is statistically different from zero is not complete without the reviewer understanding how much the effect sizes vary across studies. This is, in fact, often viewed as a prelude to the most interesting analyses in a meta-analysis. In fixed effect models, the test statistic Q provides a test of whether the amount of variation among effect sizes is due only to expected sampling variability, or equivalently, whether the effect sizes from all studies are equal. The test for homogeneity is a formal test of the hypothesis that all effect sizes, ES_i are estimating the same population parameter. The test statistic Q is given by

$$Q = \sum w_i(ES_i - \overline{ES})^2. \quad (15)$$

When all the effect sizes are estimating the same population parameter, then Q follows a χ^2 distribution with $k - 1$ degrees of freedom. When we reject the null hypothesis, Q follows a noncentral χ^2 distribution with $k - 1$ degrees of freedom and a noncentrality parameter, λ, given by

$$\lambda = \sum w_i (ES_i - \overline{ES})^2. \quad (16)$$

The prospective power of Q, the test of homogeneity, is

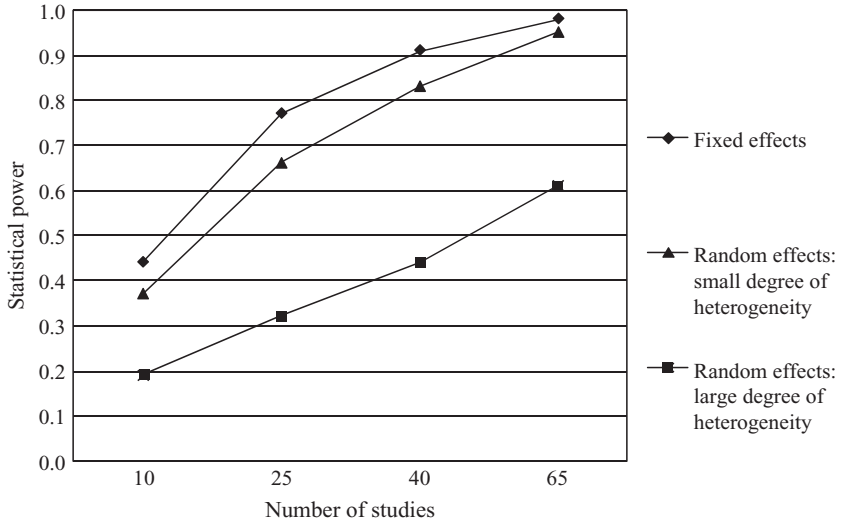

FIGURE 2. *Statistical power to detect a population effect size of* $\delta = .15$*, assuming a typical within-study sample size of 40. Large and small degrees of heterogeneity are defined using the conventions described in the article.*

$$p = 1 - F(c_\alpha \mid k - 1; \lambda), \tag{17}$$

where $F(c_\alpha \mid k - 1; \lambda)$ is the cumulative distribution function of the noncentral χ^2 distribution with $k - 1$ degrees of freedom for the noncentrality parameter λ. Note that the noncentrality parameter has the same form as Q. To compute prospective power, we need an estimate of λ, or equivalently, of the anticipated value of the homogeneity statistic, Q, and a method to estimate the cumulative distribution function of the noncentral χ^2. For λ, we can use the framework we developed in Figure 1 to provide estimates of small, moderate, and large amounts of heterogeneity. If we examine the equation for λ, we see that the quantity that is squared within the summation is a measure of the difference between each individual study's effect size and the mean of all effect sizes (i.e., it is analogous to a sum of squares in an analysis of variance [ANOVA]). This quantity is then divided by the variance, v_i, of each study's effect size. If we assume that all studies have a similar effect size variance, then the value of Q is the sum over k studies of the variance of the study effect sizes divided by the common value of v, or, equivalently, k multiplied by the ratio of between-study variance (the variance among the study effect sizes) to the within-study variance (the variance of each study's effect size). We can use our earlier convention from Figure 1 to state that a small degree of homogeneity would correspond to one third as much variability between effect sizes as within studies, or $\lambda = k(.33)$. By the same convention,

a moderate amount of variability would result in $\lambda = k(1.0)$ and a large amount to $\lambda = k(3.0)$.

To compute values for the noncentral χ^2 distribution, we need to use standard statistical packages such as statistical package for social sciences (SPSS) or Statistical Analysis System (SAS; Microsoft's Excel spreadsheet program only provides the standard central χ^2 distribution). In SPSS, the appropriate function (in the Transformations menu) is the noncentral probability density function, NPDF.CHISQ($c_\alpha \mid df, \lambda$) where c_α is the critical value for the central χ^2 with *df* is the degrees of freedom, and λ is the noncentrality parameter.

To compute the power for our example, recall that we have identified $k = 40$ studies in our review. A small amount of heterogeneity in our review corresponds to $\lambda = 40(.33) = 13.2$, a moderate amount of heterogeneity to $40(1.0) = 40$, and a large amount of heterogeneity to $40(3.0) = 120$.

Given that the 95% critical value of the χ^2 with 39 degrees of freedom is 54.57, the power in our example for detecting a small amount of between-studies heterogeneity in the population of effect sizes can be computed from the SPSS function NPDF.CHISQ, and is $p = 1 - F(54.57|39; 13.2) = 1 - 0.03 = 0.97$. In this case, we have sufficient power to detect a small amount of effect size heterogeneity in our sample of 40 studies, and we would also have similarly high power to detect moderate and high degrees of heterogeneity.

Power of tests of moderators of effect size: ANOVA models. When there is a statistically significant amount of variation among effect sizes, reviewers often examine whether characteristics of the studies are associated with the heterogeneity. Categorical moderator variables can be tested using procedures similar to one-way ANOVA models under both the fixed effects and random effects models.

As in ANOVA, reviewers test whether the group mean effect sizes, as defined by the categorical moderator variable, are estimating the same population parameter. In earlier sections, we discussed the assumptions needed to compute the power of tests of the overall mean effect under the fixed effects and random effects models. The power of the test under the fixed effects model that the group means are all equal depends on (a) the number of studies in each group, (b) the value of the group means, and (c) the variance of the group means. As seen in the prior sections, there are many ways to arrive at an estimate for the value of the group means and their associated variance. The power of the test under the random effects model requires an estimate of the between-groups variance component in addition to the assumptions listed above.

Suppose that in our example of studies on adolescent depression we are interested in whether there is a difference in the effectiveness of prevention programs when they occur in school versus when they occur outside a school setting. As in one-way ANOVA, we use an omnibus test of the hypothesis that the group mean effect sizes are equal. For p different group means, the test statistic has a χ^2

distribution with $(p - 1)$ degrees of freedom and is referred to in many meta-analysis texts as the between-group homogeneity statistic, Q_B. We compute Q_B from

$$Q_B = \sum_{i=1}^{p} \frac{1}{v_{i\bullet}} (\overline{ES_{i\bullet}} - \overline{ES})^2, \tag{18}$$

where $v_{i\bullet}$ is the variance for the ith group mean, $\overline{ES_{i\bullet}}$ is the value for the ith group mean, and $\overline{ES}$ is the overall group mean. To compute the power, we first need an estimate of the variance for the ith group mean. If we make the same assumptions as earlier in the article, then we have a variance for each study's effect size of 0.1003. If we also assume that 20 studies examine the effectiveness of depression prevention programs during school and the remaining 20 studies examine studies with out-of-school programs, we can use Equation 4 to arrive at the variance of the group means, or,

$$v_{1\bullet} = v_{2\bullet} = \frac{0.1003}{20} = 0.005.$$

We also need to estimate the squared differences between the group means and the overall mean. With some algebraic manipulation and a focused comparison involving only two groups, we can show that

$$\sum_{i=1}^{2} \frac{1}{v_{i\bullet}} \left(\overline{ES_{i\bullet}} - \overline{ES}\right)^2 = \frac{\left(\overline{ES_1} - \overline{ES_2}\right)^2}{v_1 + v_2}, \tag{19}$$

simplifying our computations. In our earlier example, an effect size of $\delta = 0.15$ was considered a meaningful difference on self-report measures. Thus, instead of making conjectures about the values of individual group means, we can state that the value of the smallest meaningful difference between the two group means is 0.15.

The power of the omnibus test that the group means are equal uses the test statistic the between-group homogeneity statistic Q_B. When the null hypothesis is true, that is, when all the population means are equal, Q_B has a χ^2 distribution with $p - 1$ degrees of freedom. When at least one of the group means differs from the other means, then Q_B has a noncentral χ^2 distribution with noncentrality parameter, λ_B, equal to

$$\lambda_B = \sum_{i=1}^{p} \frac{1}{v_{i\bullet}} (\overline{ES_{i\bullet}} - \overline{ES})^2. \tag{20}$$

The power of the omnibus test of the group mean differences is given by

$$1 - F(c_\alpha | p - 1; \lambda_B), \tag{21}$$

where $F(c_\alpha \mid p - 1;\ \lambda_B)$ is the cumulative distribution function of the noncentral χ^2 distribution with $p - 1$ degrees of freedom and noncentrality parameter λ_B. This distribution is given in SPSS as the function NCDF.CHISQ as discussed earlier. For our example of the difference between the effect size means for studies whose programs take place in school day versus outside school, we have a noncentrality parameter, λ_B, for a difference between group means of 0.15 by using Equation 19:

$$\lambda_B = \frac{(0.15)^2}{(0.005 + 0.005)} = 2.25.$$

The 95% critical value of a central χ^2 distribution with 1 degree of freedom is 3.84. We now have all the information we need to compute the power of the omnibus test of the group mean differences for this example using Equation 21:

$$1 - F(3.84 \mid 1, 2.25) = 1 - 0.68 = 0.32.$$

If there are more than two groups, Hedges and Pigott (2004) provide the power for contrasts among group mean effect sizes. Essentially, in the fixed effects case, the power of the test of a contrast assuming fixed effects is based on the *Z* test or standard normal distribution.

In the random effects case, or a mixed effects model with a categorical predictor, the procedures are similar except we also need an estimate of $\hat{\tau}^2$, the variance component for the set of studies. For example, to test the null hypothesis that the group mean effect sizes are equal in a mixed effects model, we use the test statistic

$$Q_B^* = \sum \frac{1}{v_{i\bullet}^*} \left(\overline{ES_{i\bullet}^*} - \overline{ES^*}\right)^2, \tag{22}$$

where $\sum 1/v_{i\bullet}^*$ is the sum of the inverse of the variance of the *i*th effect size in the random effects model, or $\sum_{j=1}^{m_i} 1/\left[\hat{\tau}^2 + v_{ij}\right]$. The values $\overline{ES_{i\bullet}^*}$ and $\overline{ES^*}$ are the random effects group means and the overall random effects mean, respectively. When the null hypothesis is true, that is, when all the group mean effect sizes are equal, the test statistic Q_B^* has a central χ^2 distribution with $(p - 1)$ degrees of freedom, where p is the number of means under comparison. When the null hypothesis is not correct, Q_B^* has a noncentral χ^2 distribution with $(p - 1)$ degrees of freedom and noncentrality parameter λ_B^* given by

$$\lambda_B^* = \sum \frac{1}{v_{i\bullet}^*} \left(\overline{ES_i^*} - \overline{ES^*}\right), \tag{23}$$

where $\overline{ES_{i\bullet}^*}$ and $\overline{ES^*}$ are the random effects group mean effect sizes and the overall mean effect size, respectively. The power of the omnibus test of the group means is then

$$1 - F(c_\alpha \,|\, p - 1;\, \lambda_B^*), \tag{24}$$

where $F(c_\alpha \,|\, p - 1;\, \lambda_B^*)$ is the cumulative distribution function of a noncentral χ^2 with $p - 1$ degrees of freedom and noncentrality parameter λ_B^* evaluated at the $(1 - \alpha)$ critical value c_α of a central χ^2 with $p - 1$ degrees of freedom.

To compute power for the between-groups random effects mean differences, we need a value of λ_B^*. When we have only two groups, the noncentrality parameter λ_B^* is equivalent to

$$\lambda_B^* = \frac{\left(\overline{ES_1^*} - \overline{ES_2^*}\right)^2}{\left(v_{1\bullet}^* + v_{2\bullet}^*\right)}, \tag{25}$$

where the $v_{i\bullet}^*$ are the random effects variance for the group means. One difficulty is deciding on an appropriate approximation for the random effects variances for the group means.

If we assume that each group has the same number of effect sizes, m, and the sample size within each study is equivalent (leading to a common value of the effect size variance, v), then the value of $v_{i\bullet}^*$ is equal to

$$v_{i\bullet}^* = v_{1\bullet}^* = v_{2\bullet}^* = \sum_{i=1}^{m} \left(\tau^2 + v\right) = m\left(\tau^2 + v\right), \tag{26}$$

where τ^2 is the random effects variance component within each group, and v is the common variance for each effect size. If we use the same convention we developed for the test of homogeneity in fixed effects, a small amount of heterogeneity would correspond to a ratio of τ^2 to the typical random effects variance of a study's effect size (v) of 1/3, or $\tau^2/v = 1/3$. A moderate amount of heterogeneity would correspond to $\tau^2/v = 1.0$, and a large amount to $\tau^2/v = 3$, or equivalently, $\tau^2 = 3v$. Thus, for a small amount of heterogeneity, $v_{i\bullet}^*$ would equal

$$v_{i\bullet}^* = m\left(\frac{v}{3} + v\right) = \frac{4mv}{3}.$$

Using the same logic, a moderate amount of heterogeneity would correspond to $v_{i\bullet}^* = 2mv$ and a large amount of heterogeneity to $v_{i\bullet}^* = 4mv$.

In our example, we have $m = 20$ effect sizes within each of the two groups, with a common value of variance for each effect size of $v = 0.1003$. Thus, if we wish to find the power for detecting a group mean difference of 0.15, then for a small amount of heterogeneity, the noncentrality parameter, λ_B^*, is equal to

$$\lambda_B^* = \frac{(0.15)^2}{2\left(\frac{4*20*0.1003}{3}\right)} = \frac{3 * (0.15)^2}{8 * 20 * 0.1003} = \frac{0.068}{16.05} = 0.004.$$

Using Equation 24, the power of the test of the difference between the two random effects means is given by

$$1 - F(3.84|\,1,\,0.004) = 1 - 0.95 = 0.05.$$

Given our assumptions, the power calculation suggests that only 5% of the time would correctly reject a false null hypothesis regarding the presence of a moderating variable. Of course, the situation gets worse with larger degrees of between-study heterogeneity.

Power of tests for regression models of effect size. Effect size regression modeling, or meta-regression, allows the reviewer to test models of effect size that include both continuous and categorical predictors. These models can contain either fixed or random predictors or a mix of both. Typically, a reviewer will compute tests of individual regression coefficients or all coefficients in the model simultaneously and also will conduct tests of goodness of fit to assess model adequacy. To obtain the results of these tests, and by extension of their statistical power, the reviewer needs to know the values of the individual predictors or, if these are not available, the full covariance matrix of the predictors with the effect sizes and the effect size variances. Thus, reviewers will find it difficult to compute the power of particular meta-regression tests if it is not possible to obtain the full covariance matrix of the predictors and effect sizes in the model. In our experience, this is usually not the case, and as such, we will not provide a demonstration in this article. Hedges and Pigott (2004) provide an example of tests for meta-regression using a data set that includes a set of predictors. The reviewer could compute power for a meta-regression model if the full covariance matrix is obtained (perhaps by contacting study authors) and if the reviewer has decided a priori what size of regression coefficient will represent an important finding.

Integrating cluster randomized trials. For many research questions posed in educational contexts, both individual and cluster randomized trials may have been conducted and therefore both are eligible for inclusion in a systematic review. Cluster randomized trials can be problematic in this regard because they are often analyzed at the level of the individual (or more generally, at the level of the microunit). For example, if 10 classrooms were randomly assigned to deliver either a prevention program or a placebo program, the data might be analyzed at the level of the student. However, doing so ignores the strong possibility that responses within classrooms will be more similar to each other than they will be to responses in other classrooms. The likely dependencies in the data cause, among other things, estimated standard errors to be too small. If this is not taken into account, then a study would get too much weight in a meta-analysis and similarly, the study could contribute to a false sense of precision when investigating power.

One relatively simple approach to both problems involves computing an "effective sample size," that is, an estimate of the amount of independent information contributed by a study. Computing the effective sample size involves knowing the number of clusters in the study, the average cluster size, and the interclass correlation ρ; together these are used to compute a "design effect" (Kish, 1965). The intraclass correlation coefficient expresses the degree of similarity of observations within a cluster (it is equivalent to the average correlation between any two randomly sampled units in the same randomly sampled cluster). Unfortunately, ρ is rarely reported in research studies, so sources outside the study usually need to be consulted. Fortunately, for some types of data, compendia are available that will help reviewers with this task (e.g., Hedges & Hedberg, 2006, for academic achievement outcomes; Verma & Lee, 1996, for health), and more may become available as a function of the increasing attention paid to the problem of violating the assumption of independence.

The design effect is computed as

$$1 + (m - 1)\rho, \tag{27}$$

where m is the average cluster size (e.g., the average number of students in a classroom) and ρ is the intraclass correlation. The effective sample size is computed as the total sample size of microunits (e.g., students) divided by the design effect. When $\rho = 1$, the effective sample size is equal to the number of clusters; when $\rho = 0$, the effective sample size is equal to the number of microunits.

Assume that reviewers believe that of the 40 studies they expect to find, they anticipate that approximately 10%, or 4, will be cluster trials and that the rest are trials in which individuals were assigned to conditions. To account for the dependencies in the four cluster trials, they will need to find a credible external source or guess at ρ. Assume that a credible external source cites $\rho = .10$ as a "typical" value for self-report measures of depression among adolescents and that a scoping review suggested that an average of 3 classrooms per condition, with 22 students each, were used in the studies. From Equation 27, we know that the design effect is $1 + [(22 - 1) \times .10)] = 3.1$. The effective sample size is the total number of students divided by the design effect, here $132/3.1 = 42.58$. Thus, by these assumptions the typical cluster trial in this area would overstate the number of amount of independent information by about 68% (i.e., 1 − the inverse of the design effect). Therefore, if these studies were used to inform the guesses about the likely sample size for the original prospective power analysis, then those analyses may have been overly optimistic. Because it was estimated that cluster trials would make up about 10% of the total studies, and that the sample sizes (relative to the effective sample sizes) were overestimated by about 68%, it follows that adjusting down the overall sample size by about 6.8% might provide a better estimate of prospective power. In this case, the effect on estimated power would be slight (from .91 to .89 for fixed effects power to detect $\delta = .15$), but in other cases the impact might be nontrivial, for example, with a larger intraclass

correlation, a higher proportion of cluster trials among the included studies, or with larger within-cluster sample sizes.

How many studies are needed for sufficient power? Reviewers considering initiating a meta-analysis might be interested in determining the number of studies that need to meet inclusion criteria for the meta-analysis to have sufficient power to detect the effect of interest. This might just be the easiest part of the entire systematic review, as answering this question involves only a little algebra that follows from the power analysis. As was done in the previous power analyses, reviewers will need to postulate a typical within-study sample size and will also need to either (a) determine the smallest important effect size given the research context or (b) make an educated guess about the effect size that is likely to be found. As noted earlier, they will also need to specify a Type I error rate and whether they will be conducting a one- or two-tailed test. The difference in the current situation is that instead of estimating the number of relevant studies likely to be identified, the reviewers will need to make a decision about what constitutes "sufficient" statistical power.

Assume that, based on Cohen (1988), the reviewers have decided that "sufficient" power means that they have an 80% chance of correctly rejecting a false null hypothesis and further assume that all other values are estimated to be the same as they were in the previous examples (i.e., typical within-study sample size is 20 per cell and the smallest important effect is $\delta = 0.15$). From the fixed effects example we know that $v = 0.1003$. From Equation 6, we also know that $\lambda = \overline{\text{ES}}/\sqrt{\text{v}_\bullet}$ and from Equation 3 that $\text{v}_\bullet = \text{v}/k$. We also know that the standard normal cumulative distribution function that yields a value of 0.2 is $-.842$ (again, using any statistical textbook with a z table or the Microsoft Excel function "=normsinv"), which means that $1.64 - (-0.842)$ or 2.482 is the value for which we need to solve (i.e., 2.482 is the value of λ that delivers power of approximately 0.80). Finding the number of studies necessary to have power of .80 is simply a matter of solving Equation 6 for k:

$$\frac{.15}{\sqrt{.1003/k}} = 2.482, \quad k = 27.4.$$

In other words, approximately 27 studies are needed to meet inclusion criteria for the significance test of the overall average effect size to have power of approximately 0.80, given a fixed effects analysis and the other assumptions made by the reviewers.

Computing the number of studies, given a random effects analysis is similar, except that the parameter v is replaced with v*, which in the random effects example was 0.2006. Because we know from Equation 12 that $\text{v}^*_\bullet = \text{v}^*/k$, we can substitute and solve Equation 11 for k:

$$\frac{.15}{\sqrt{.2006/k}} = 2.482, \quad k = 54.9.$$

Alternatively, if fixed effects power has already been computed, this quantity is simply the number of studies identified under fixed effects assumptions times (v^*/v). Either way, for this example, approximately 55 studies are needed to meet inclusion criteria for the significance test of the overall average effect size to have power of approximately 0.80, given a random effects analysis and the other assumptions made by the reviewers.

Reviewers occasionally find themselves in the position of finding a very large number of studies that meet inclusion criteria. In these cases, one option that might be adopted for the sake of efficiency is to code and analyze a random sample of studies. The procedures we outline above can be used to determine how many studies ought to be randomly sampled and coded. In this case, assuming the researchers were interested in obtaining power of about 0.80, they would need to code about 27 or 55 studies, depending on whether they planned on conducting a fixed or a random effects analysis.

Retrospective Statistical Power

For many researchers who rely on tests of statistical significance for support of their hypotheses, the first time they seriously consider the issue of low statistical power is when they have conducted a statistical test that has failed to reject the null hypothesis. These researchers discover the need to assess the likelihood that low statistical power has led to an incorrect statistical conclusion, often referred to as a Type II error (i.e., the failure to correctly reject a false null hypothesis), relative to the likelihood that their substantive hypothesis was wrong. As in retrospective (or post hoc) power computations for primary analyses, there are two ways to carry out this analysis in meta-analyses.

A typical (if uninformative) way of carrying out this analysis would be to use all the observed values to compute power. As an example, assume a review team found 11 studies, with an average within-study sample size of 19, and an overall average effect size of $d = 0.18$. A fixed effects analysis would not result in the rejection of the null hypothesis, $d = 0.18$, $p = .10$ one-tailed for our hypothetical data. Retrospective power can be computed using procedures outlined for prospective power analysis, except that instead of making assumptions about the size of the effect that they are interested in for theoretical or clinical reasons, reviewers will often input the *observed* values from the studies they have collected. That is, the researchers will assume that the population parameter (δ) they were trying to detect is equal to the sample statistic that they observed (d) and will input into the analysis the observed number of studies (11), the observed average within-study sample size (19), and the Type I error rate that they used.

For our current example, power given the observed values was approximately 0.37, suggesting that the analysis did not have enough power to detect the observed effect.

The reason that we have labeled this an uninformative exercise is because the observation that the analysis did not have enough power to detect the effect is true *by definition* when all the observed values from a nonstatistically significant test form the basis of the power analysis. As such, when done in this manner, the retrospective power analysis does not add information to the analysis already conducted. Note that this is equally true for primary analyses (Hoenig & Heisey, 2001).

A somewhat better approach for assessing power retrospectively is to use the observed number of studies and within-study sample size but to choose an effect size based on considerations relevant to the context of the research question. For example, the researchers could base retrospective power computations on the smallest effect size they believe is meaningful, or on some other value that is meaningful in the context of the research question, rather than on the mean observed effect. In the example above, assume the reviewers were working on an intervention meant to prevent adolescent depression. Assume further that the most popular self-report scale for assessing literacy in adolescents is scaled to have a mean of 100 and a standard deviation of 15. This information could be used to make guesses about the smallest true effect size that would be important. Here, assume that the reviewers decided that any true effect <1.5 points on the literacy scale would not be important. From Equation 1, we can see that the researchers' smallest important standardized mean difference effect size is $1.5/15 = 0.10$. The retrospective power to detect this (fixed) effect given 11 studies and a within-study sample size of 19 is .18, one-tailed. An example of how reviewers might report this result is

> The overall average effect size was not statistically significant, $d = 0.18$, $p = .10$ one-tailed. However, given the number of studies we found and the average within-study sample size in those studies, power to detect what we believe to be the smallest important effect was low (approximately .20). As such, the failure to reject the null hypothesis should be interpreted with caution. Perhaps the most reasonable conclusion to be drawn from this result is that we currently do not have enough information to judge adequately whether this intervention has a meaningful effect on the literacy of participants.

Alternatively, one could compute power for a range of plausible population effect size values (e.g., Rossi, 1990). For example, researchers might compute power for the upper and lower values of a confidence interval, as well as some intermediate points. Regardless, relative to the strategy of using all observed values (i.e., number of included studies, average within-study sample size, effect size estimate, Type I error rate) to compute power, the strategies that base power

computations on substantive considerations or on a range of empirically plausible effect sizes add useful information beyond that given by the probability value from the statistical test.

The use of retrospective power analysis is especially important in the case of moderator variables (see Hedges & Pigott, 2004). Often, systematic reviewers are interested in the impact of categorical moderators on effect size. In these cases, the researchers will compare groups in a meta-analysis, fail to find a significant difference, and conclude (perhaps erroneously) that the moderator has no effect on the treatment. Moderator tests are particularly likely to be underpowered, and thus it is important to put the failure to reach statistical significance in perspective using retrospective power analysis of the sort just described.

Statistical Power as a Guide to Updating a Systematic Review

One final benefit of statistical power assessments concerns the ability to use these as a guide to determining whether an update to a review will be beneficial. In the section "How many studies are needed for sufficient power?" above, we outlined how to use prospective power analysis to determine the number of studies that would be needed to have a given level of power, conditional on certain assumptions made by the reviewers. Those procedures can be used after a meta-analysis has been conducted and power found to be low, the main difference being that, such as with our suggestions for retrospective power, the reviewers will use the observed parameters (with the exception of the effect size, which should be chosen by reviewers because it represents an important effect) to compute the number of studies that are needed for a certain level of power. The difference between the number of studies needed for the desired power and the number of studies that met inclusion criteria represents the number of additional studies that need to be conducted and/or located before updating the review.

Confidence Interval Estimation

Some scholars (e.g., Levin & Robinson, 2000; Robinson & Levin, 1997) have argued that given a nonsignificant inferential statistical test, the most reasonable estimate of the population effect is zero; indeed, this is a common misperception (Rosenthal & Rubin, 1994). Based on what we have said so far, we hope that readers realize that we believe this to be a misguided advice. Probability values suggest how likely one is to observe a given (or more extreme) effect size under a true null hypothesis. Another way of thinking about these values is that they convey how confident one can be that the direction of a relationship has been correctly identified. They do not suggest the confidence that can be placed in the effect size estimate itself. As such, especially given conditions of low statistical power, the absence of statistical significance should be interpreted as "more information is required" rather than "there is no effect." We would like to point

out the opposite as well; to regard statistically significant probability values as reflecting important effect sizes is also mistaken. Statistical significance does not reflect meaningfulness. In a large sample, it could reflect a substantively trivial effect. If one of the virtues of meta-analysis is that it increases the sample size so that it increases the likelihood that the null hypothesis can be rejected, it may also be a vice, if statistical significance is misinterpreted as important, meaningful, or precise.

We have come to think of this problem as a case of mistaken identity with a missing accomplice, in which the probability value is mistakenly viewed as an indicator of the magnitude of the relationship and in which the confidence interval estimate (as an index of precision of the effect size) has disappeared entirely. We have a strong preference for separating the tasks of effect size estimation and interval estimation (i.e., estimating the most likely value and the plausible range of values separately) and we argue that the best approach to interpretation of meta-analytic results is one that uses both the point estimate and the confidence interval, either in addition to the *p* value or in place of it.

The effect size with its related confidence interval provides all the information that is provided by a test of statistical significance and additionally provides other information that is of great use in interpreting the results of the meta-analysis (see Cumming & Maillardet, 2006). Separating the estimate of effect size from the precision of this estimate focuses attention on the substantive impact of the intervention. The lower bound of the confidence interval tells us whether the result is statistically significant (it is not if the confidence interval overlaps with zero) and also something about the minimum likely size of the effect. The upper bound of the confidence interval allows us to assess the "best case" for the intervention's impact, something that the *p* value says nothing about. Finally, the width of the confidence interval (irrespective of whether it includes zero) provides the full range of plausible parameters that are consistent with the results of the meta-analysis and thus provides information about how precisely we can answer the question "Does the intervention work?"

In the hypothetical example above, the 95% confidence interval ranged from a low of -0.09 to a high of $+0.45$. As such, it seems reasonable to suggest that the intervention is unlikely to harm participants and may confer benefits. Indeed, if the smallest important effect size is $\delta = 0.10$, then the confidence interval suggests the intervention could plausibly lead to benefits that are substantially larger than that. Although the result was not statistically significant, the wide confidence interval indicates that the intervention's effect was measured too imprecisely to interpret that result with confidence.

A useful real-life example may be found in a review of training interventions for foster parents. Turner, MacDonald, and Dennis (2005) found two studies that examined the impact of foster parent training on the psychological health of foster children. The weighted average of these two studies was $d = +0.13$, with a 95% random effects confidence interval that ranged from $d = -0.71$ to

$d = +0.96$. Because the confidence interval overlapped with zero, we know that the statistical test for the average effect size was not statistically significant at $\alpha = .05$ (in fact, for this test $p > .90$). The standard textbook interpretation of this test is "We could not reject the null hypothesis." Readers following Levin and Robinson (2000) might be tempted to conclude that the best estimate of the population parameter is zero. However, we think that the most critical information lies in the confidence interval. It suggests that the plausible range of values includes negative effects that, if real, would be evidence of clinically meaningful harms and positive effects that, if real, would be evidence of clinically meaningful benefits. This finding is far more important than the point estimate associated with that confidence interval or the fact that these results were not statistically significant. Like our hypothetical example above, it seems that the most reasonable conclusion to be drawn from this analysis is not that the intervention does not work but rather that we really have very little information about whether it works, a fact highlighted by the very wide confidence interval.

It should also be noted that confidence intervals provide meaningful information even when the statistical analysis results in rejection of the null hypothesis. For example, assume that a group of reviewers believe that the smallest important effect size for their research question is $\delta = 0.25$. They conduct a study, and their results are statistically significant, $d = 0.251, p = .049$. The 95% confidence interval for this analysis would suggest a low value of approximately $d = 0.001$ and a high value of approximately $d = 0.501$. Most reviewers would be happy to have a statistically significant result and an observed effect size that was large enough to be considered important. These positive developments would probably lead to a claim that the intervention "works." However, it should be clear from the confidence interval that slightly less than half of the plausible values of the true intervention effect include values that the researchers would not consider important. This information is much harder to see in a power analysis, if indeed a retrospective power analysis were carried out. Here, the power to detect a true population effect of $\delta = .25$ would be approximately .50, assuming all the other parameters that went into the computation of the confidence intervals remained constant.

Options for Reviews That Do Not Result in a Synthesis

Should a Synthesis be Carried Out if Power Is Low?

When the computed prospective power for a planned meta-analysis is low, the reviewers may question whether it is worthwhile for them to invest the time and effort needed to conduct the systematic review. When we are asked this question, we generally answer "Yes, it is still worthwhile." It should be clear by now that prospective power analysis depends on assumptions; these assumptions include the number of studies that will meet inclusion criteria, the typical sample size within studies, the size of the effect being investigated, and, for random effects

analyses, the size of the between-study variance component. Of course, these assumptions could be wrong. For example, a thorough literature search that goes beyond the usual electronic databases or that is creative in generating search terms or identifying gray literature can reveal a surprising number of eligible studies, at times considerably more than the reviewers expected to find. Or, the population effect size might be larger than the effect size on which the power calculations were based. If either or both of these happened, a prospective power analysis would underestimate actual power. An additional benefit arising from doing the meta-analysis is that it allows the reviewers to compute and interpret confidence intervals and as such adds information beyond what is revealed in the individual studies.

After collecting studies, reviewers may again consider whether they should statistically combine studies when power is low (e.g., if few studies were found that met inclusion criteria). What should reviewers do if they decide that the studies constituting their evidence base cannot be combined? We have already suggested that the primary conclusion ought to be that more evidence is needed. In addition, it will often be helpful for reviewers to describe the studies (and their findings) that met their inclusion criteria. However, before discussing these options, we first outline strategies that reviewers have used, which we believe to be fundamentally flawed. It is our firm belief that given the need for some kind of synthesis, all the available alternatives are worse than meta-analysis, in that they are likely to be based on less defensible assumptions and on less transparent processes. As such, systematic reviewers who are not content, or do not have the option, to simply state "There is not enough evidence to draw conclusions about this relationship" are likelier to arrive at a valid conclusion if they use meta-analysis than if they use other techniques such as vote counting or doing a "cognitive algebra" synthesis.

How Not to Synthesize Studies

Sometimes reviewers reach narrative conclusions regarding the effect of an independent variable even when a quantitative synthesis has not been carried out. This practice in the research literature takes the form of scholars beginning with the assertion that "the studies were too heterogeneous to combine" (perhaps without ever operationally defining what "too heterogeneous" means) and ending with assertions about the overall nature of the effect (e.g., "Taken as a whole, the studies seem to indicate that . . . "). This is a seriously limited inferential technique, for several reasons.

First, in effect, a statement that studies are too different to combine implies that the studies are not all estimating the same population parameter, which is an essential assumption underlying fixed effects meta-analysis. However, if the assertion that the studies are "too heterogeneous to combine" is taken seriously, it precludes both a quantitative and a qualitative summary in most circumstances.

Furthermore, the overall summary arises from the unique "cognitive algebra" of the reviewers (a term we borrow from social psychology to refer to the idiosyncratic and unstated rules individuals use to arrive at a conclusion). The main problem with a cognitive algebra synthesis is that there is little reason to believe that neither other scholars will know what dimensions (and how they were weighted) were relevant to the decision (i.e., the decision-making process will not be transparent) nor will they likely be able to reach a similar conclusion under similar circumstances (i.e., the decision will not be reproducible).

An only slightly better approach (better due to its transparency) might be to synthesize studies using the results of the individual statistical tests. For instance, if there are two studies and both effects are in the same direction and statistically significant, the reviewers could claim that the intervention is effective. More problematically (and more realistically given typical power conditions) if one or few studies are statistically significant and others fail to reach significance, or if effects were in different directions, the evidence would be labeled as "mixed." Or, if all studies fail to reject the null hypothesis of no effect then the reviewers would claim that there is no effect (regardless of the patterns of the directions of the effect).

By now the limits of this approach, known as vote counting, should be clear. Although it is relatively transparent, vote counting has properties that seriously limit its validity as an inferential technique. First, even if both studies are estimating the same population parameter in practice, it is relatively unlikely that both studies will result in a rejection of the null hypothesis. For example, if both studies are conducted with average power (Cohen, 1962; Sedlmeier & Gigerenzer, 1989) of .50, then the probability that both will correctly reject a false null hypothesis only .25. Even under conditions of high power (.80), the probability of both studies rejecting a false null hypothesis is only .64. Both studies would have to be conducted under conditions of extraordinary statistical power (.975) for there to be a 95% chance that they would both correctly result in a rejection of a false null hypothesis. As such, requiring that both studies reject the null hypothesis of no difference is not likely to be a valid approach to synthesis.

Furthermore, vote counting ignores the possibility of Type I (e.g., arising from the conduct of "too many" statistical tests) and Type II errors (e.g., due to low statistical power). A related and alarming point is the Hedges and Olkin (1985) demonstration that in conditions of low-to-moderate statistical power in primary studies (a situation common in the social sciences; Sedlmeier & Gigerenzer, 1989), vote counting actually has *less* statistical power, the *more* evidence accumulates. As such, as a stand-alone inferential technique, vote counting based on the statistical significance of the findings has little to recommend it. In some cases, however (e.g., very impoverished data reporting), vote counting based on the direction of the observed effects might be a reasonable technique (see Bushman, 1994).

Another option might be to require the effect sizes from both studies to be roughly similar. This approach has the virtue of not relying on statistical

significance testing, but it does rely on judgment about whether effect sizes are similar. It is unclear what rules ought to be applied to that judgment, and highly likely that the rules will be unstated, and that different reviewers will apply different rules. Furthermore, simply comparing the effect sizes visually does not take advantage of the information provided by the precision with which those effects are estimated.

Meta-analysis, however, provides a method for taking advantage of the relevant information comprising the statistical significance tests in the studies (i.e., effect sizes and their precision), avoids the problems associated with using the statistical conclusions arising from individual tests, and does so in a transparent and replicable way. In this sense, the answer to the question "How many studies do you need to do a meta-analysis?" is "two." Not because it is ideal but rather because given the need for a conclusion (e.g., an administrator who needs to pick a program), it is a better analysis strategy than the alternatives.

When it is legitimate not to synthesize studies? We do believe that there are times in which summarizing the results of multiple studies is not appropriate. For example, Cooper (2003) suggests that a meta-analysis of two studies will likely only be informative if the studies are direct (or "statistical") replications of one another. The combination of very few studies with very different characteristics makes any kind of synthesis untenable in most cases. Furthermore, with very few studies, parameter estimation (e.g., the point estimate, the random effects variance component) will likely be poor, rendering conclusions that are highly uncertain.

Ideally, reviewers will have given some thought to the question of the conditions under which they will synthesize the data during the planning of the systematic review. They could, for example, consider the outcomes they are likely to find and whether these ought to be combined in a single analysis. Similar consideration can be given to whether it makes sense to collapse across certain variations in the intervention, sample, research design, and so on. We recognize that this is a difficult exercise that should be considered a tentative guide to the systematic review. However, we believe that it represents a significant source of protection against data-driven decisions that might serve to protect the interests of the reviewers. Formal planning for a systematic review is relatively common in medical research. In addition, both the Cochrane Collaboration and the Campbell Collaboration, two organizations that support the production of systematic reviews, require detailed protocols before reviewers can start collecting data. Regardless of whether the review was based on a formal protocol, we believe it is important for reviewers to be as specific as possible about the reasons for not synthesizing data and to be clear about the timing of that decision (i.e., whether it is based on decisions outlined in a protocol or on decisions made after the data had been collected).

Calls for More Evidence

Many reviewers find themselves in the position of Oliver Twist ("Please, sir, I want some more"; Dickens, 1838/1997, chap. 2). Of course, instead of thin gruel they are asking for additional research to be conducted. Probably the most helpful way for reviewers to frame this request is for them to be explicit about what they see as the limitations in the evidence as it currently exists. For example, there may be a very small number of studies on the topic, or the studies that have been conducted are rife with plausible threats to their validity, or there may be important gaps in the sampled participants, and so on. The more prescriptive reviewers can be about the characteristics of needed studies, the more helpful their recommendations are likely to be to future scholars and funders. Therefore, we suggest that reviewers be very specific about the limitations they perceive in a literature and how these might be addressed in future studies. These recommendations could include, for example, the identification of important theoretical variables that are understudied, an elaboration of the weaknesses in existing research design and how these undermine claims about the existence of a relationship, and an examination of the likely external validity of existing studies and how this might be augmented in subsequent studies.

Narrative Summaries of Studies

In addition to providing specific recommendations about the next generation of research, reviewers should consider methods for describing the current generation of research. Importantly, they will need to do so in a way that overcomes some of the limitations of traditional narrative review procedures, especially the habitual overreliance on statistical significance testing. One way to do this is to include text and/or tables for each study. These should describe the nature of the intervention, sample, outcome measures, and research design. In essence, by providing this level of detail, the reviewers are defending their decision not to summarize the results of the individual studies. As such, the most important considerations in that decision ought to be addressed in the narrative and/or tables.

Furthermore, reviewers should provide their readers with an estimate of the magnitude of the intervention's effect as well as confidence intervals for the effect sizes. Reviewers could also provide the results of any associated statistical significance tests. Consider, for example, the following descriptions of three research studies presented in Tables 3 and 4. The descriptions in Table 3 follow a fairly traditional narrative review procedure and express results solely in terms of their statistical significance. The descriptions in Table 4 have the same information as in Table 3 and also provide both effect sizes and 95% confidence intervals for each study. Table 4 contains all the information from Table 3 but by adding the effect sizes and confidence intervals, contextualizes the results of the statistical significance tests more effectively. It is clear from the confidence intervals that wide range of population values for the effect of prevention

TABLE 3
Traditional Narrative Description of the Characteristics and Results of Individual Studies

Study	Narrative Description
Mays et al. (1999)	100 students were randomly assigned to participate in a prevention program or to be on a wait-list. Program effects were assessed via the Beck Depression Inventory. There was no significant difference between the groups.
Mantle et al. (2000)	20 students attending a local prevention program were compared to 20 "matched controls" on a previously published scale tapping attributional style. It was unclear how matching was implemented. There were no significant differences between the groups.
Snider et al. (2001)	60 students attending a prevention program at a local Boys & Girls Club were compared to 60 controls. The groups were matched on multiple measures of psychological functioning. Results revealed a positive effect, with students attending the prevention program performing better on a locally developed measure of depressive symptoms than students not attending the program.

Note: Studies referred to in table are fictitious.

TABLE 4
Alternate Narrative Description of the Characteristics and Results of Individual Studies

Study	Narrative Description	Effect Size ± 95% CI
Mays et al. (1999)	100 students were randomly assigned to participate in a prevention program or to be on a wait-list. Program effects were assessed via the Beck Depression Inventory.	+0.15 ± 0.34
Mantle et al. (2000)	20 students attending a local prevention program were compared to 20 "matched controls" on a previously published scale tapping attributional style. It was unclear how matching was implemented.	+0.30 ± 0.62
Snider et al. (2001)	60 students attending a prevention program at a local Boys & Girls Club were compared to 60 controls. The groups were matched on multiple measures of psychological functioning. The groups were compared on a locally developed measure of depressive symptoms.	+0.37 ± 0.36

Note: The effect size is expressed in standardized mean difference units. Studies referred to in table are fictitious. CI = confidence interval.

programs are plausible, even for the study which resulted in a rejection of the null hypothesis.

Table 5 presents yet another suggestion for how to portray the critical features of study design and results. Here, several important elements of the studies

TABLE 5
Tabled Presentation of the Characteristics and Results of Individual Studies

Study (Year)	Treatment *n*	Comparison *n*	Assignment Mechanism	Matching Used?	Outcome Variable	Effect Size	95% CI Lower Limit	95% CI Upper Limit
Mays et al. (1999)	50	50	Random	No	Beck Depression Inventory	+0.15	−0.19	+0.49
Mantle et al. (2000)	20	20	Nonrandom	Yes, matching variables unknown	Published attributional style scale	+0.30	−0.32	+0.92
Snider et al. (2001)	60	60	Nonrandom	Yes, multiple measures of psychological functioning	Ad hoc measure of depressive symptoms	+0.37	+0.01	+0.73

Note: The effect size is expressed in standardized mean difference units. Studies referred to in table are fictitious. CI = confidence interval; *n* = sample size.

(e.g., research design, type of intervention) and their results (e.g., effect size and confidence intervals) are displayed in column format. We believe that this design is the best way to present vital information to readers and may be the only workable option when more than a few studies have met inclusion criteria.

Discussion

Our hope in this article was to continue the evolution of the practice of systematic reviewing. To this end, we started by outlining fixed and random effect power analysis. Like prospective statistical power analysis for primary studies, prospective power analysis for meta-analysis requires assumptions about parameters that are unknown before undertaking the review. We provided some suggestions for thinking about these, in particular for the random effects variance component. Specifically, we argued that one way of thinking about the likely degree of between-study heterogeneity is the extent to which this heterogeneity is "built into" the analysis through the choices made by the reviewers. We also show how this information can be used in conjunction with guidelines from medical research regarding the degree of between-study heterogeneity to estimate the between-studies variance component. Next, we discussed retrospective power analysis, suggested that this analysis is often carried out in an uninformative manner, and then showed how this analysis can be made more informative by computing power based on some important effect size (rather than on the observed effect size). We then discussed the value of confidence intervals, showed how they could be used in addition to or instead of retrospective power analysis, and also demonstrated that confidence intervals can convey more information in some situations than power analyses alone.

Finally, we took up the question "How many studies do you need to do a meta-analysis?" and showed that, given the need for a conclusion, the answer is "two studies," because all other synthesis techniques are less transparent and/or less likely to be valid. For systematic reviewers who choose not to conduct a quantitative synthesis, we provide suggestions for both highlighting the current limitations in the research base and for displaying the characteristics and results of studies that were found to meet criteria for inclusion in the review.

References

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*, 19–32.

Birge, R. T. (1932). The calculation of errors by the method of least squares. *Physical Review*, *40*, 207–227.

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 193–214). New York: Russell Sage.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NY: Lawrence Erlbaum.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods*, *8*, 243–253.

Cooper, H. (2003). Psychological Bulletin: Editorial. *Psychological Bulletin*, *129*, 3–9.

Cooper, H. (2008). The search for meaningful ways to express the effects of interventions. *Child Development Perspectives*, *2*, 181–186.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, *11*, 217–227.

Dickens, C. (1838/1997). *Oliver twist; or, the Parish boy's progress*. Seattle, WA: The World Wide School. Retrieved January 5, 2009, from http://www.worldwideschool.org/library/books/lit/charlesdickens/OliverTwistOrtheParishBoysProgress/legalese.html

Hedges, L. V., & Hedberg, E. C. (2006). Intraclass correlations for planning group randomized experiments in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87.

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

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*, 203–217.

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., & Thompson, S. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558.

Higgins, J., Thompson, S., Deeks, J., & Altman, D. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical Research Edition)*, *327*, 557–560.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*, 19–24.

Howell, D. C. (1992). *Statistical methods for psychology* (3rd ed.). Boston: PWS-Kent Publishing Company.

Kish, L. (1965). *Survey sampling.* New York: John Wiley.

Levin, J. R., & Robinson, D. H. (2000). Rejoinder: Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher*, *29*, 34–36.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance. *Educational Researcher*, *26*, 21–29.

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science*, *5*, 329–334.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*, 646–656.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

Torgerson, C. (2003). *Systematic reviews and meta-analysis.* London: Continuum International Publishing Group.

Turner, W., MacDonald, G. M., & Dennis, J. A. (2005). Behavioural and cognitive behavioural training interventions for assisting foster carers in the management of difficult behaviour. *Cochrane Database of Systematic Reviews*. Art. No.: CD003760. DOI: 10.1002/14651858.CD003760.pub3.

Verma, V., & Lee, T. (1996). An analysis of sampling errors for demographic and health surveys. *International Statistical Review*, *64*, 265–294.

Authors

JEFFREY C. VALENTINE is Associate Professor in the department of Educational and Counseling Psychology, University of Louisville; e-mail: jeff.valentine@louisville.edu. As coeditor of the *Handbook of Research Synthesis and Meta-Analysis* (2nd ed.), his interests are in strengthening validity and utility of systematic reviews, and in applying rigorous systematic review techniques to address practical problems in the social and behavioral sciences.

THERESE D. PIGOTT is Associate Professor in the research methodology program in the School of Education, Loyola University Chicago. Her research interests are statistical methods for meta-analysis, the use of hierarchical linear models, and methods for handling missing data in statistical analysis.

HANNAH R. ROTHSTEIN is Professor of Management at Baruch College and the Graduate Center of the City University of New York, where she coordinates the doctoral specialization in Organizational Behavior, chairs the Baruch IRB, and teaches courses in research methods. She is a coauthor of software for meta-analysis (*Comprehensive Meta-Analysis*) and for power analysis (*Power and Precision*), coeditor of *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments, and coauthor of Introduction to Meta-Analysis*.

Manuscript received January 12, 2009
Revision received January 28, 2009
Accepted May 7, 2009