Psychological Methods
2001, Vol. 6, No. 3, 203–217

1082-989X/01/$5.00 DOI: 10.1037//1082-989X.6.3.203

The Power of Statistical Tests in Meta-Analysis

Larry V. Hedges
University of Chicago

Therese D. Pigott
Loyola University of Chicago

Calculations of the power of statistical tests are important in planning research studies (including meta-analyses) and in interpreting situations in which a result has not proven to be statistically significant. The authors describe procedures to compute statistical power of fixed- and random-effects tests of the mean effect size, tests for heterogeneity (or variation) of effect size parameters across studies, and tests for contrasts among effect sizes of different studies. Examples are given using 2 published meta-analyses. The examples illustrate that statistical power is not always high in meta-analysis.

The use of quantitative methods to summarize the results of several empirical research studies, or meta-analysis, is now widespread in psychology, medicine, and the social sciences. Meta-analysis involves describing the results of each study using a numerical index (an estimate of effect size such as a correlation coefficient, a standardized mean difference, or an odds ratio) and then combining these estimates across studies to obtain a summary.

Although inference procedures for meta-analysis have been available for well over a decade, there is little work on the calculation of the power of statistical tests in meta-analysis. However, power calculations are always part of sound statistical planning (Cohen, 1977). Moreover, power calculations are often a required component of research grant proposals in primary research, and the requirement of providing some estimate of statistical power is increasingly an issue in evaluating research synthesis projects as well. Although meta-analyses with large numbers of studies investigating even medium-sized effects may have quite powerful tests, meta-analyses of smaller numbers of studies and meta-analyses in areas in which effects are expected to be small do not necessarily have very powerful statistical tests.

The purpose of this article is to provide procedures to calculate the statistical power of the statistical tests most frequently used in meta-analysis. These include fixed- and random-effects tests on the mean effect size across studies (Hedges & Vevea, 1998), tests for heterogeneity of effect size parameters across studies (Hedges, 1982a), tests for contrasts among effect sizes (Hedges, 1982b; Rosenthal & Rubin, 1982), and tests for differences among groups of studies (Hedges, 1982b). Before discussing the details of power computations in meta-analysis, we first argue for the importance of conducting informative power analyses before investing resources in a quantitative research synthesis.

As others have argued (Cohen, 1977; Kraemer & Thiemann, 1987; Lipsey, 1990), power analysis provides important information prior to beginning any research study. Several investigators have developed software to assist in these computations (Bornstein, 2000; Elashoff, 2000; O'Brien, 1998). Power analyses are used to plan studies by ensuring that the power of the statistical tests used will be adequate for the smallest effect size deemed to be of practical significance in a given context. The same reasoning holds true for meta-analysis. As those who have conducted any comprehensive research review know, searching and obtaining a representative sample of studies on a given topic requires a large investment of time, money, and energy. Electronic search capabilities have eased some of the burden in recent years, but reviewers still must search the unpublished literature carefully, often performing hand searches of journals and personally calling active researchers in the field. No researcher wants to begin a meta-analysis project if there is little chance that the findings will prove useful. Power analyses conducted prior to a meta-

Larry V. Hedges, Department of Education, University of Chicago; Therese D. Pigott, School of Education, Loyola University of Chicago.

Correspondence concerning this article should be addressed to Larry V. Hedges, Department of Education, University of Chicago, 5835 South Kimbark Avenue, Chicago, Illinois 60637. Electronic mail may be sent to hedge@src.uchicago.edu.

analysis can provide the reviewer with the likelihood of finding a statistically significant result given the anticipated size of the overall effect, the number of studies included in a review, and the typical sample size within studies.

In power analysis in primary studies, the power of statistical tests depends on parameters that the researcher does not know before conducting the study. Similarly, in meta-analysis, the power of statistical tests depends on information about studies that a reviewer typically does not know before conducting the literature search. Waiting to do the power analysis after the studies have been collected goes against the purpose of power computations; if one has sufficient power, then the review will probably yield statistically significant findings. How then does a reviewer perform power calculations if the information needed about the sample of studies is unknown?

The question of how to conduct a power analysis without definitive prior information is common to the planning of all experiments. However, few statisticians would argue that not doing a power analysis is better than computing power with approximate information. The statistical power of tests in (fixed-effects) primary analyses of research typically depends on effect size, the level of statistical significance, and the sample size. Power analysis in primary research therefore requires setting a threshold for the smallest effect size considered of substantive importance and the sample size for a given level of significance. In principle, sample size and level of statistical significance are under the control of the investigator, but there are often practical limitations on the former and limitations of convention on the latter.

Statistical power of tests in (fixed-effects) meta-analysis also typically depends on effect size, the level of statistical significance, and the sample size. However, in the case of meta-analysis, the sample size has two components: the number of studies and the within-study sample size (which is usually functionally related to the variance of a study's effect size estimate). Power analysis in meta-analysis therefore requires guesses about the effect size and both components of the sample size (the number of studies and the within-study sample size) for a given level of significance.

In random- or mixed-model statistical analyses in primary research, the statistical power also depends on other parameters that may be characterized in various ways (usually as variance components or intraclass correlations); see, for example, Snijders and Bosker (1993), Maxwell and Delaney (1990, pp. 568–575), or Diggle, Liang, and Zeger (1994, pp. 28–29). Similarly, in random- or mixed-model meta-analysis, the statistical power also depends on another parameter (the between-studies variance component). Although the need for the value of an additional parameter complicates these power analyses (compared with fixed-effects analyses), it is no greater complication in meta-analysis than in the corresponding primary analysis.

How can the needed information about effect size and sample sizes be obtained? In power analysis in primary research, previous studies are often suggested as a source of estimates of parameters needed for power analyses. In meta-analysis, previous reviews in the area, using either qualitative or quantitative methods, are the analogous source of information. For example, Rind, Tromovitch, and Bauserman (1998) recently published a meta-analysis on the effects of child sexual abuse in college students. As they noted in their literature review, several reviews have appeared in the published literature on the effects of child sexual abuse. We use this case as an example later in this article to illustrate that Rind et al. might have used information from these reviews to construct estimates of the power of statistical tests in their meta-analysis. In using this example, we acknowledge the controversies surrounding the findings of the Rind et al. meta-analysis. In particular, the findings from Rind et al. cannot be generalized to all victims of child sexual abuse given that the meta-analysis was limited to nonclinical samples attending college. As we demonstrate below, the study might not have been carried out if power analyses had been conducted prior to the meta-analysis. Lipsey and Wilson (1993) provided a comprehensive list of meta-analyses in psychology and education; from their list, one can imagine that many important issues have been reviewed at least qualitatively.

In using previous reviews, it is important to recognize that new meta-analyses may have a different (often narrower) focus than previous reviews. For example, Rind et al. (1998) used their meta-analysis not only to add new research findings to prior reviews but also to focus more closely on a subset of the studies included in this previous work, specifically those involving college students. A reviewer might also conduct a new review using different assumptions from prior reviews, such as using random-effects models instead of fixed-effects models. Finally, a reviewer might intend to use different inclusion rules for a new review than those used in prior reviews (e.g., including only the most recent studies). Any of these changes might lead to a situation in which the number of studies in the new review is not necessarily larger than in previous reviews, and thus a statistically significant finding in a previous review is not necessarily an indication of adequate power in a later meta-analysis conducted for a different purpose.

If there are no previous studies in an area, primary researchers often use data from pilot studies to estimate power. The analogous process in meta-analysis is a pilot review. A search of an existing database such as PsycINFO or ERIC could provide a count of the possible numbers of studies on a given topic. A reviewer must be cautious, however, about the estimate of the number of studies. For example, a review of the effects of child sexual abuse by Neumann, Houskamp, Pollock, and Briere (1996) identified 488 studies but used only 38 of those studies in a quantitative review. Guesses about the size of the effect and typical within-study sample size can often be derived from study abstracts. Note that one need not actually examine all of the studies or abstracts. Examination of a sample could be used to obtain estimates of the number of studies with sufficient information to enter the meta-analysis and the within-study sample sizes.

When neither a previous study nor a pilot study is available, primary researchers are often advised to rely on similar studies or expert opinion (the judgment of the investigator or experts in the field). Such an approach is also possible in meta-analysis, but in either case it is difficult to know a priori whether this information and any power computations based on it are very accurate.

All statisticians who perform power analyses on primary studies understand that we rarely have all the information we need before we conduct the study. In the case of meta-analysis, researchers rarely attempt a comprehensive review of the literature without knowing that some number of studies exists. Meta-analysts, like primary researchers, must base their power analyses on prior work and rely on expert knowledge of the area of interest.

Statistical Inference in Meta-Analysis

In this article, we assume that there are effect size estimates from k independent studies. We denote the population effect size (effect size parameter) in the ith study by θ_i and its estimate (the sample effect size estimate) by T_i:

$$\begin{array}{cccc} 1 & \theta_1 & T_1 & v_1 \\ 2 & \theta_2 & T_2 & v_2 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ k & \theta_k & T_k & v_k. \end{array}$$

We assume that the T_i are normally distributed about the corresponding θ_i with known variance v_i. That is, we assume that

$$T_i \sim N(\theta_i, v_i)\ i = 1, \ldots, k. \qquad (1)$$

This assumption is nearly exactly true for effect sizes such as Fisher's z-transformed correlation coefficient and standardized mean differences transformed by the Hedges–Olkin variance-stabilizing transformation (Hedges & Olkin, 1985). However, for effect sizes such as the untransformed standardized mean difference or correlation coefficient, or the log-odds ratio, the results are not exact but remain true as large-sample approximations.

Two somewhat different statistical models have been developed for inference about effect size data from a collection of studies, called the *fixed-effects model* and the *random-effects model,* respectively (Hedges & Vevea, 1998). Fixed-effects models treat the effect size parameters as fixed but unknown constants to be estimated and usually (but not necessarily) are used in conjunction with assumptions about the homogeneity of effect parameters (Hedges, 1982a; Rosenthal & Rubin, 1982). Random-effects models treat the effect size parameters as if they were a random sample from a population of effect parameters and estimate hyperparameters (usually just the mean and variance) describing this population of effect parameters (DerSimonian & Laird, 1986; Hedges, 1983; Schmidt & Hunter, 1977).

In the sections that follow, we consider computation of the statistical power of tests on the mean effect size and homogeneity of effect sizes using first fixed-effects and then random-effects statistical methods. Then we provide a method to compute the power of tests of contrasts among effect sizes. These calculations are exact when the conditional variances of the effect size parameters are known exactly but only approximate (based on large-sample theory) when the v_i are not known exactly. Finally we consider the accuracy of the power calculations when the conditional variances are not known exactly.

Statistical Inference in Fixed-Effects Meta-Analysis

If a series of k studies can reasonably be expected to share a common effect size θ, it is natural to estimate θ by pooling estimates from each of the studies. If the sample sizes of the studies differ, then the estimates from the larger studies will usually be more precise than the estimates from the smaller studies. In meta-analysis, we use a weighted estimator with weights inversely proportional to the precision or variance in each study. The optimal weights w are given by

$$w_i = 1/v_i. \tag{2}$$

Thus the weighted mean that minimizes the variance can be written as

$$\bar{T}. = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}. \tag{3}$$

Note that $\bar{T}.$ is also the maximum-likelihood estimator of θ under this model. Note also that this estimate, or any estimate, of a common θ may be misleading if the θ_i vary substantially.

The sampling variance $v.$ of $\bar{T}.$ is simply the reciprocal of the sum of the weights, namely,

$$v. = \frac{1}{\sum_{i=1}^{k} w_i}, \tag{4}$$

and the standard error $SE(\bar{T}.)$ of $\bar{T}.$ is just the square root of $v.$, that is, $SE(\bar{T}.) = \sqrt{v.}$. Because $T_1, \ldots, T_k$ are normally distributed, it follows that $\bar{T}.$ is also normally distributed.

Tests for the Mean Effect Size

If $T_1, \ldots, T_k$ estimate the same underlying effect size $\theta_1 = \ldots = \theta_k = \theta$, then $\bar{T}.$ estimates θ and a 100α percent significance test of the null hypothesis that $\theta = \theta_0$ could be obtained using the statistic

$$Z = (\bar{T}. - \theta_0)/\sqrt{v.}, \tag{5}$$

which has the standard normal distribution when $\theta = \theta_0$. The one-tailed test rejects the null hypothesis whenever $Z > c_\alpha$, where c_α is the 100 $(1 - \alpha)$ percentile of the standard normal distribution (e.g., for $\alpha = .05$, $c_\alpha = 1.645$). The two-tailed test rejects the null hypothesis whenever $|Z| > c_{\alpha/2}$.

The statistic Z in the test given above has the standard normal distribution when $\theta = \theta_0$. If $\theta \neq \theta_0$, Z has a normal distribution with mean

$$\lambda = (\theta - \theta_0)/\sqrt{v.} \tag{6}$$

and variance of 1.

Because the one-tailed test at significance level alpha rejects the null hypothesis if $Z > c_\alpha$, the power of the one-tailed test that $\theta = \theta_0$ is given by

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

where $\Phi(x)$ is the standard normal cumulative distribution function.

The computation of the power of the two-tailed test is only slightly more complicated. The two-tailed test at significance level alpha rejects the null hypothesis if $|Z| > c_{\alpha/2}$, that is, if $Z > c_{\alpha/2}$ or if $Z < -c_{\alpha/2}$. Therefore the power of the one-tailed test that $\theta = \theta_0$ is given by

$$p = 1 - \Phi(c_{\alpha/2} - \lambda) + \Phi(-c_{\alpha/2} - \lambda). \tag{8}$$

If the v_i values are thought to be approximately equal, then it follows that $v.$ is approximately v/k, where v is the common value of the v_i. However, it should be noted that if the v_i values are not identical, and $\bar{v}$ is the average of the $\bar{v}_i$, $\bar{v}/k$ will be strictly larger than $v.$, and using $\bar{v}/k$ in place of $v.$ in power calculations will result in an underestimate of the statistical power.

Example: Standardized mean differences. Neumann, Houskamp, Pollock, and Briere (1996) conducted a review of the effects of childhood sexual abuse in women prior to Rind et al.'s (1998) synthesis. The Neumann et al. study included both clinical and nonclinical samples and focused only on the effects of abuse on women. Though Rind et al. focused on nonclinical samples, and specifically included the effects on male victims, the Neumann et al. study can provide guesses about the magnitude of the effect size, the variance of the effect sizes, and the number of studies for the computation of power.

For each study in the review, Neumann et al. (1996) computed the standardized mean difference between women who reported sexual victimization during childhood (CSA positive) and women who did not report abuse during childhood (CSA negative) on some measure of psychological functioning. Hedges and Olkin (1985) defined the standardized mean difference, d_i, as

$$d_i = (\overline{Y}_{\mathrm{P}_i} - \overline{Y}_{\mathrm{N}_i})/s_i,$$

where $\overline{Y}_{\mathrm{P}_i}$ and $\overline{Y}_{\mathrm{N}_i}$ are the sample means of the CSA-positive and CSA-negative women for study i and s_i is the pooled sample standard deviation for study i. The estimated variance of d_i is given by

$$v_i = \frac{n_{\mathrm{P}_i} + n_{\mathrm{N}_i}}{n_{\mathrm{P}_i} n_{\mathrm{N}_i}} + \frac{d_i^2}{2(n_{\mathrm{P}_i} + n_{\mathrm{N}_i})},$$

where n_{P_i} and n_{N_i} are the sample sizes of CSA-positive and CSA-negative women in the ith study, respectively.

To calculate the power for the Rind et al. (1998) study, we need to posit reasonable values for $v.$ and k. Table 1 presents a summary taken from Neumann et al.'s (1996) article. As can be seen in Table 1, Neumann et al. found an average effect size for nonclinical samples of $T. = 0.32$, meaning that CSA-positive women on average scored a little more than one fourth of a standard deviation higher on measures of psychological symptoms than CSA-negative women. The 95% confidence interval for the average effect size was given by Neumann et al. as (0.27, 0.37). As a conservative estimate, we could use the value $T. = 0.20$ as a minimally significant finding, meaning that we would want the ability to detect a difference between CSA-positive and CSA-negative adults that is less than one-fourth of a standard deviation.

Table 1 also provides an estimate of $v.$, the variance of the overall effect size, which depends on the variance of the effect sizes in each study. The 95% confidence interval for the weighted mean effect size could provide one estimate of $v.$ given that the lower limit is given by Hedges and Olkin (1985) as $\delta_{\mathrm{L}} = d. - c_{\alpha/2}\sqrt{v.}$. Alternatively, we could posit a reasonable value for the within-study sample size. Neumann et al. (1996) reported that in their restricted sample of 37 studies (including studies using clinical samples), 5 studies had within-study sample sizes less than 50, and the remaining studies had sample sizes ranging from 51 to 500. Neumann et al. also reported that their original sample of 38 studies represents 2,774 CSA-positive women and 8,388 CSA-negative women, a ratio of 1:3. If we took $N = 48$ for the within-study sample size, with $n_{\mathrm{P}} = 12$ and $n_{\mathrm{N}} = 36$, we have an estimate of the common value of v_i of

$$v_i = \frac{12 + 36}{(12)(36)} + \frac{(0.2)^2}{(2)(12)(36)} = 0.111.$$

The value of the variance of the weighted mean effect size is given by $v. = v/k = 0.111/18 = 0.0062$, where k, the number of studies with nonclinical samples included in Neumann et al.'s review, is 18. For a fixed-effects analysis, $\lambda = (0.20 - 0)/\sqrt{0.0062} = 0.20/0.079 = 2.53$. The power for a one-tailed test when $\alpha = .05$ is given by Equation 7, where $p = 1 - \Phi(1.64 - 2.53) = 1 - \Phi(-0.89) = 1 - 0.19 = .81$ and $c_{.05} = 1.64$ is the 95th percentile of the standard normal distribution. The power of the two-tailed test at significance level $\alpha/2 = .025$ is given by inserting the value of λ into Equation 8 to obtain $p = 1 - \Phi(1.96 - 2.53) + \Phi(-1.96 - 2.53) = 1 - \Phi(-0.57) + \Phi(-4.49) = 1 - 0.28 + 0.0 = .72$. In this case, we have a reasonable amount of power to detect a difference of 0.20 with 18 studies, each having a minimum of 48 subjects with a ratio of 1:3 for CSA-positive versus CSA-negative adults.

Table 1
Summary of Results From Neumann, Houskamp, Pollock, and Briere (1996)

Study moderator	k	$d.$	95% CI	Q_B
Source of recruitment				9.40*
Clinical	17	0.50	0.40–0.61	
Nonclinical	18	0.32	0.27–0.37	
Mixed	2	0.43	0.20–0.67	
Locus of abuse				1.73
Intrafamilial	8	0.47	0.29–0.69	
Mixed	29	0.35	0.30–0.39	

Note. From "The Long-Term Sequelae of Childhood Sexual Abuse in Women," by D. A. Neumann, B. M. Houskamp, V. E. Pollock, and J. Briere, 1996, *Child Maltreatment, 1*, p. 10. Copyright 1996 by Sage Publications. Reprinted with permission. k = the number of independent studies; $d.$ = the weighted mean effect size; CI = confidence interval; Q_B = the homogeneity test.
$^*p < .05$.

Example: Correlation coefficients. Rind et al. (1998) also discussed a number of qualitative reviews in their synthesis, including one by Kendall-Tackett, Williams, and Finkelhor (1993). Though this review did not use meta-analysis, the authors provided enough information to enable us to make good guesses about possible values of $v.$ and $k.$ for the power analyses. For the measure of effect size, Kendall-Tackett et al. provided preliminary calculations from a subset of studies using η^2, a measure of the proportion of variance accounted for on a measure of psychological functioning. The value of η^2 captures both linear and nonlinear components of the relationship between two variables. Kendall-Tackett et al.

found a range of η^2 values on many different psychological symptoms of .01 to .77.

As discussed by Cohen (1977, p. 283), we could use the value

$$\sqrt{\eta^2} = \eta$$

as our effect size so that it has the same scale as r, the correlation coefficient. However, when η is applied to a study with two groups, this approximation only holds true when there are equal numbers of subjects per group. In our example, r would approximate η if the studies have equal numbers of abused and nonabused participants, which holds true in some of the studies. If this condition is not true, the analyst should transform η^2 to d as illustrated by Cohen (1977). We use r as an approximation to η here to illustrate power computation with r.

Kendall-Tackett et al. (1993) found a range of values of η from .10 to .88. We can posit a value of $r = .10$ as a conservative value for the effect size in these studies. As described in Hedges and Olkin (1985), the distribution of the sample correlation coefficient depends on the unknown value of the population correlation and is nonnormal. Transforming sample correlation coefficients to Fisher's z where

$$z = \tfrac{1}{2} \log [(1 + r)/(1 - r)]$$

normalizes the distribution. In this case, a value of $r = .10$ gives a Fisher's z value of 0.10.

We now need to estimate $v_{\cdot}$, the variance of the weighted mean for Fisher's z. The variance of Fisher's z is $v_i = 1/(n_i - 3)$, which depends on sample sizes within studies, n_i. Kendall-Tackett et al. (1993) reported a range of sample sizes from 8 to 369 but also stated that the majority of studies used sample sizes of between 25 and 50. We can take $n = 25$ as a conservative estimate of within-study sample size. Thus, we have an estimate of $v_i = 1/(25 - 3) = 1/22 = 0.045$.

To obtain an estimate of $v_{\cdot}$, we need an estimate of k, the number of studies. Although Kendall-Tackett et al. (1993) identified 45 studies for review, not all studies used nonclinical samples. Table 2 provides a summary of numbers of studies identified by Kendall-Tackett et al. Depending on the psychological symptom, a range of 6 to 38 studies use nonclinical samples, with a mean of about 17 studies. We could take 10 as a conservative guess of the number of studies focusing on a particular symptom.

Table 2
Summary of Studies Identified by Kendall-Tackett, Williams, and Finkelhor (1993)

	Total no. of studies	
Symptom	Nonclinical	Clinical
Anxiety	14	3
Fear	6	3
Posttraumatic stress disorder	8	2
Depression	38	10
Poor self-esteem	11	
Somatic complaints	16	7
Mental illness	15	10
Aggression	24	12
Sexualized behavior	25	8
School–learning problems	13	3
Behavior problems	31	7
Self-destructive behavior	9	
Composite symptoms	21	6

Note. From "Impact of Sexual Abuse on Children: A Review and Synthesis of Recent Empirical Studies," by K. A. Kendall-Tackett, L. M. Williams, and D. Finkelhor, 1993, *Psychological Bulletin, 113*, p. 166. Copyright 1993 by the American Psychological Association. Reprinted with permission.

If we estimate that we will gather at least 10 studies ($k = 10$) with a sample size of 25 and a Fisher's z transformation equal to 0.10, our estimate of $v_{\cdot}$ can be computed as $v_{\cdot} = v/k$, or $[1/(25 - 3)]/10 = (1/22)/10 = 0.0045$. The power of the one-tailed test at $\alpha = .05$ for $\zeta = 0$ requires the computation of $\lambda = (0.10 - 0)/\sqrt{0.0045} = 1.49$. The power to reject the hypothesis that $\zeta = 0$ at $\alpha = .05$ is given by inserting λ into Equation 7 to obtain $p = 1 - \Phi(1.64 - 1.49) = 1 - \Phi(0.15) = 1 - 0.56 = .44$. The two-tailed power of the test at significance level $\alpha = .05$ ($\alpha/2 = .025$) is given by inserting λ into Equation 8 to obtain $p = 1 - \Phi(1.96 - 1.49) + \Phi(-1.96 - 1.49) = 1 - \Phi(0.47) + \Phi(-3.45) = 1 - 0.68 + 0.00 = .32$. Given our assumptions of $r = .10$, $k = 10$, and common within-study sample size of 25, we have little power to detect a correlation of .10 between child sexual abuse exposure and adult outcomes.

Testing for Heterogeneity of Effect Size Parameters

Before pooling the estimates of effect size from a series of k studies, it is important to determine whether the studies can reasonably be described as sharing a common effect size. A statistical test for the homogeneity of population effect sizes is formally a test of the hypothesis

$$H_0\colon \theta_1 = \theta_2 = \ldots = \theta_k$$

versus the alternative that at least one of the effect sizes θ_i differs from the remainder.

An exact small-sample test of H_0 (which is also the likelihood ratio test of this hypothesis) is based on the statistic

$$Q = \sum_{i=1}^{k} w_i(T_i - \overline{T}.)^2, \qquad (9)$$

where $\overline{T}.$ is the weighted estimator of effect size given in Equation 3. The test statistic Q is the sum of squares of the T_i about the weighted mean $\overline{T}.$, where the ith square is weighted by the reciprocal of the variance of T_i. Because $(T_i - \overline{T}.)^2$ can be seen as a (crude) estimate of between-studies variation, each term of Q can also be interpreted as a ratio of between-studies to within-study variances given that $w_i = 1/v_i$.

If all k studies have the same population effect size (i.e., if H_0 is true), then the test statistic Q has a chi-square distribution with $k - 1$ degrees of freedom. Therefore, if the obtained value of Q exceeds the $100(1 - \alpha)$ percent critical value of the chi-square distribution with $k - 1$ degrees of freedom, we reject the hypothesis that the θ_i values are equal.

The test statistic Q is given in Equation 9, which has the chi-square distribution with $k - 1$ degrees of freedom when the null hypothesis of homogeneity ($\theta_1 = \ldots = \theta_k$) is true. When the null hypothesis of homogeneity is not true, that is, when some of the effect sizes differ, then Q has a noncentral chi-square distribution with $k - 1$ degrees of freedom and noncentrality parameter λ given by

$$\lambda = \sum_{i=1}^{k} w_i(\theta_i - \overline{\theta}.)^2, \qquad (10)$$

where $\overline{\theta}.$ is the weighted mean of $\theta_1, \ldots, \theta_k$ given by

$$\overline{\theta}. = \frac{\sum_{i=1}^{k} w_i\theta_i}{\sum_{i=1}^{k} w_i}. \qquad (11)$$

The power of the test based on Q at significance level alpha is therefore

$$p = 1 - F(c_\alpha | k - 1; \lambda), \qquad (12)$$

where $F(x | \upsilon; \lambda)$ is the cumulative distribution function of the noncentral chi-square with υ degrees of freedom and noncentrality parameter λ and where c_α is the $100(1 - \alpha)$ percent point of the central chi-square distribution. This distribution is tabulated and widely available in statistical software.

A rough approximation to the noncentral chi-square distribution using the central chi-square distribution was given by Patnaik (1949). This approximation gives the distribution of Q as approximately a constant

$$a = 1 + \lambda/(k - 1 + \lambda)$$

times a central chi-square distribution with

$$\upsilon = (k - 1) + \lambda^2/[(k - 1) + 2\lambda]$$

degrees of freedom. The power is approximately

$$1 - F(c_\alpha/a | \upsilon; 0), \qquad (13)$$

where $F(x|\upsilon; 0)$ is the cumulative distribution function of the central chi-square distribution with υ degrees of freedom.

Conventions for heterogeneity. Although Equations 12 and 13 provide expressions for the power of the heterogeneity test, they are not useful unless a value of λ can be computed, which in turn depends on the values of the individual effect size parameters $\theta_1, \ldots, \theta_k$, which may be difficult to guess in the preliminary stages of a meta-analytic study. An alternative procedure for developing plausible values of λ when the v_i values are identical (and might be used if they are nearly so) is to treat λ as $(k - 1)/v$ times the "variance" of the θ_i, and thus λ is $(k - 1)$ times the ratio of the between-studies "variance" to the within-studies variance. Past experience in an area might suggest plausible values for this ratio. For example, Schmidt (1992) examined many meta-analyses in psychology and found that the ratio of between-studies variance to within-studies variance rarely exceeds one and that 0.33 is a more typical value (corresponding to Schmidt and Hunter's, 1977, 75% rule—the average conditional variance is 75% of the total variance of the estimates). Therefore, one might adopt the convention that $\lambda = .33(k - 1) = (k - 1)/3$ is a small degree of heterogeneity, $\lambda = .67(k - 1) = 2(k - 1)/3$ is a medium degree of heterogeneity, and $\lambda = (k - 1)$ is a large degree of heterogeneity.

Example: Power of the homogeneity test. Suppose we only have the information provided in the Kendall-Tackett et al. (1993) review. In order to compute an approximate power of the homogeneity test, we must make a number of assumptions. Our earlier example used conservative estimates for the common within-study sample size ($n = 25$) and for the number of studies ($k = 10$). Because we do not know the values of the θ_i (in this case, estimates of the individual correlations from each study), we have to assume a specific degree of heterogeneity among the θ_i to make the statistical power computation. Here we calculate three possible values of λ, the noncentrality parameter, based on whether we assume there is a small degree of heterogeneity, medium heterogeneity, or high heterogeneity. That is, $\lambda = (k - 1) * .33$ (small heterogeneity), $\lambda = (k - 1) * .67$ (medium heterogeneity), or $\lambda = (k - 1) * 1.0$ (large heterogeneity), which correspond to $\lambda = (10 - 1) * .33 = 2.97$, $\lambda = (10 - 1) * .67 = 6.03$, and $\lambda = (10 - 1) * 1.0 = 9.0$.

The statistical program SPSS (1999) provides the noncentral distribution function of the chi-square in the transformation menu as NCDF.CHISQ (*q, df, nc*) where q is the c_α, 100 $(1 - \alpha)$ percent point of the central chi-square distribution; *df* is the degrees of freedom or $k - 1$; and *nc* is the noncentrality parameter, or λ. In SAS (SAS Institute, 1990), the function PROBCHI (*x, df, nc*) gives the cumulative distribution of the chi-square distribution where x is equal to c_α as described above; *df* is the degrees of freedom; and *nc* is the noncentrality parameter, λ. Assuming $k = 10$, we have a value for $c_{.05}$ of 16.92, the 95% point of the chi-square distribution with $10 - 1 = 9$ degrees of freedom. For a small amount of heterogeneity, the power of the homogeneity test given in Equation 12 is $p = 1 - F(16.92|10 - 1; 2.97) = 1 - 0.83 = .17$. A moderate amount of heterogeneity yields the power as $p = 1 - F(16.92|9; 6.03) = 1 - 0.66 = .34$. With a large amount of heterogeneity, the power is $p = 1 - F(16.92|9; 9.0) = 1 - 0.49 = .51$. Given our conservative guesses about parameters of the studies, we would not expect much power to detect differences among studies in their estimates of the overall correlation between child sexual abuse and psychological outcomes.

If we have $k = 18$ studies as estimated by the Neumann et al. (1996) review, we have our estimated values for λ equal to $(18 - 1)(.33) = 5.61$ for a small amount of heterogeneity, $(18 - 1)(.67) = 11.39$ for a medium amount of heterogeneity, and $(18 - 1)(1.0) = 17$ for a large amount of heterogeneity. The value of the central chi-square with $18 - 1 = 17$ degrees of freedom at $c_{.05}$ is 27.59. The power for a small amount of heterogeneity is given in Equation 12 as $p = 1 - F(27.59|17; 5.61) = 1 - 0.77 = .23$. For a medium amount of heterogeneity, the power is $p = F(27.59|17; 11.39) = 1 - 0.50 = 0.50$. A large amount of heterogeneity gives a power value of $p = 1 - F(27.59|17; 17) = 1 - 0.28 = .72$. With a larger number of studies, we have more power to detect significant differences among effect sizes, though only if we expect a large amount of heterogeneity.

When tabulations of the noncentral chi-square are not available, Patnaik's (1949) approximation is computed from Equation 13 by estimating the auxiliary constants a and υ. With a small amount of heterogeneity and $k = 10$ studies, $a = 1 + 2.97/(10 - 1 + 2.97) = 1.25$, $\upsilon = (10 - 1) + (2.97)^2/[(10 - 1) + (2 * 2.97)] = 9.59$, and the power of the homogeneity test is $p = 1 - F(16.92/1.25|9.59; 0) = 1 - 0.83 = .17$. With a moderate amount of heterogeneity, $a = 1 + 6.03/(10 - 1 + 6.03) = 1.40$, $\upsilon = (10 - 1) + (6.03)^2/[(10 - 1) + (2 * 6.03)] = 10.73$, and the power of the homogeneity test is $p = 1 - F(16.92/1.4|10.73; 0) = 1 - 0.66 = 0.34$. With a large amount of heterogeneity, $a = 1 + 9.0/(10 - 1 + 9.0) = 1.50$, $\upsilon = (10 - 1) + (9.0)^2/[(10 - 1) + (2 * 9.0)] = 12.0$, and power of the homogeneity test is $p = 1 - F(16.92/1.4|10.73; 0) = 1 - 0.49 = .51$. All three values match those given by the exact computations.

Note that knowing the exact values of the sample sizes for the computation of the v_i does not assist in the computation of power unless one has a guess for the value of the between-studies variance. As can be seen in Equation 10, the noncentrality parameter λ is the ratio of the between-studies variance to the within-study variance. Thus, the noncentrality parameter λ can be calculated exactly if the reviewer has estimates of the between-studies variance as well as the values of the v_i. Without knowledge of the between-studies variance, the reviewer must hypothesize various plausible values of the ratio of between-studies to within-study variance as we have done in the example above.

Statistical Inference in Random-Effects Meta-Analysis

In this section we describe procedures for estimating the mean μ of the effect size distribution underlying the results of a series of studies using an analysis based on the random-effects model. There are obvious similarities between estimating a common underlying effect size by taking the mean of the estimates and estimating the mean of the effect size distribution. In both procedures, the pooled estimate is usually computed by taking the weighted mean across studies of the sample effect size estimates, and it is not unusual for either of these estimates to be called the *average* *effect size*. However, it is important to note that the quantity we are estimating (the mean of the effect size distribution) in random-effects models does not have exactly the same interpretation as the one we are estimating (the single or average underlying effect size) in fixed-effects models. In the case of random-effects models, for example, some individual effect size parameters may be negative even though μ is positive. That corresponds to the substantive idea that some realizations of the treatment may actually be harmful even if the average effect of the treatment μ is beneficial.

The Variance of Estimates of Effect Size

In the fixed-effects model, the effect sizes θ_i were fixed, but unknown, constants. Under this assumption the variance of T_i is simply v_i. In the random-effects model, the θ_i are not fixed but are themselves treated as random and have a distribution of their own. Therefore, it is necessary to distinguish between the variance of T_i assuming a fixed θ_i and the variance of T_i incorporating the variance of θ as well. The former is the *conditional sampling variance* of T_i, and the latter is the *unconditional sampling variance* of T_i.

It is convenient to decompose the observed effect size estimate into fixed and random components

$$T_i = \theta_i + \epsilon_i = \mu + \xi_i + \epsilon_i, \qquad (14)$$

where ϵ_i is a sampling error of T_i as an estimate of θ_i, and θ_i can itself be decomposed into the mean μ of the population from which the θ values are sampled and the error ξ_i of θ_i as an estimate of μ. In this decomposition, only μ is fixed, and we assume both ξ_i and the ϵ_i values are random with expected value zero. The variance of ϵ_i is v_i, the conditional sampling variance of T_i, which is known. The variance of the population from which $\xi_1, \ldots, \xi_k$ are sampled is τ^2. Equivalently, we might say that τ^2 is the variance of the population from which the study-specific effect parameters $\theta_1, \ldots, \theta_k$ are sampled. Frequently, τ^2 is called the *between-studies variance component.*

Because the effect size θ_i is a value obtained from a distribution of potential θ_i values, the unconditional sampling variance of T_i involves τ^2. A direct argument shows that this sampling variance is

$$v_i^* = v_i + \tau^2. \qquad (15)$$

Methods of estimation for random-effects models have been suggested in different meta-analytic contexts by DerSimonian and Laird (1986), Hedges (1983), and Schmidt and Hunter (1977). They make use of the method of moments to estimate the between-studies variance component and are analogous to the methods often used to estimate variance components in analyses of variance (ANOVAs) of balanced designs.

Estimating the Between-Studies Variance Component

Estimation of the between-studies variance component τ^2 uses the same principles as estimation of the variance components in ANOVA. One estimate of τ^2 is

$$\hat{\tau}^2 = \left[\begin{array}{ll} \dfrac{Q-(k-1)}{c} & \text{if } Q \geq k-1 \\ 0 & \text{if } Q < k-1 \end{array}\right., \qquad (16)$$

where c is given by

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}, \qquad (17)$$

the w_i values are the weights given in Equation 2 used in the fixed-effects analysis, and Q is given by Equation 9. Estimates of τ^2 are set to 0 when $Q - (k - 1)$ yields a negative value, because τ^2, by definition, cannot be negative.

Estimating the Mean Effect Size

The logic of using weighting is the same in random-effects procedures as it is in fixed-effects procedures, but the choice of weights differs somewhat because random-effects models include in their definition of *variance* a component of variance τ^2 associated with between-studies differences in effect parameters, which fixed-effects models do not. That is, the total variance v_i^* for the *i*th effect size estimate T_i is defined by $v_i^* = \tau^2 + v_i$. Because the additional component of variance is the same for all studies, it both increases the total variance of each effect size estimate and tends to make the total variances of the studies (the v_i^* values) more equal than the sampling error variances (v_i values).

Because the true value of τ^2 is rarely known, we usually substitute an estimate of this variance component such as that given in Equation 16 into Equation

15 in place of τ^2 (DerSimonian & Laird, 1986; Hedges & Olkin, 1985). This yields

$$\bar{T}_{\cdot} = \frac{\sum_{i=1}^{k} w_i^* \bar{T}_i^*}{\sum_{i=1}^{k} w_i^*}, \tag{18}$$

where the weight w_i^* is an estimated optimal weight that is the reciprocal of an estimate of the total variance of $\bar{T}_i$ given by

$$w_i^* = 1/(v_i^*) = 1/(v_i + \hat{\tau}^2). \tag{19}$$

Here we use the asterisk to distinguish the weights, means, and variances in the random-effects procedure from the corresponding quantities in the fixed-effects procedure.

The sampling variance $v_{\cdot}^*$ of the random-effects estimate (of the mean of the effect size distribution) $\bar{T}_{\cdot}^*$ is given by the reciprocal of the sum of the random-effects weights, that is

$$v_{\cdot}^* = \frac{1}{\sum_{i=1}^{k} w_i^*}. \tag{20}$$

The standard error $SE(\bar{T}_{\cdot}^*)$ of the mean effect estimate $\bar{T}_{\cdot}^*$ is just the square root of its sampling variance, that is, $SE(\bar{T}_{\cdot}^*) = \sqrt{v_{\cdot}^*}$. Note that whenever the between-studies variance component (estimate) $\hat{\tau}^2$ is greater than 0, the standard error $\sqrt{v_{\cdot}^*}$ of the mean estimated using the random-effects procedure will be larger than $\sqrt{v_{\cdot}}$, the standard error of the mean estimated using the fixed-effects procedure. If $\hat{\tau}^2 = 0$, the standard errors (and the mean estimates) of the random- and fixed-effects procedures will be identical.

Tests for the Mean of the Effect Size Distribution

If the random effects are approximately normally distributed, the weighted mean $\bar{T}_{\cdot}^*$ is approximately normally distributed about the mean effect size parameter μ that it estimates. As in the fixed-effects case, the fact that this mean is normally distributed with the variance given in Equation 20 leads to straightforward procedures for constructing tests of hypotheses about the mean effect size. An approximate significance test of whether the mean effect μ differs from a predefined constant μ_0 (e.g., a test of whether $\mu - \mu_0 = 0$) is accomplished by testing the null hypothesis

$$H_0\colon \mu_{\cdot} = \mu_0,$$

using the statistic

$$Z^* = \frac{\bar{T}_{\cdot}^* - \mu_0}{\sqrt{v_{\cdot}^*}}, \tag{21}$$

which has the standard normal distribution when $\mu_{\cdot} = \mu_0$. The one-tailed test consists of rejecting H_0 at level alpha if $Z^* > c_\alpha$, where c_α is the 100 $(1 - \alpha)$ percent point of the standard normal distribution (e.g., $c_\alpha = 1.645$ for $\alpha = .05$).

The test statistic Z^* has the standard normal distribution when $\mu_{\cdot} = \mu_0$, but if $\mu_{\cdot} \neq \mu_0$, Z^* has a normal distribution with mean

$$\lambda^* = (\mu_{\cdot} - \mu_0)/\sqrt{v_{\cdot}^*} \tag{22}$$

and variance 1.

Because the one-tailed test at significance level alpha rejects the null hypothesis if $Z^* > c_\alpha$, the power of the one-tailed test that $\mu_{\cdot} = \mu_0$ is given by

$$p = 1 - \Phi(c_\alpha - \lambda^*), \tag{23}$$

where $\Phi(x)$ is the standard normal cumulative distribution function.

The computation of the power of the two-tailed test is only slightly more complicated. The two-tailed test at significance level alpha rejects the null hypothesis if $|Z^*| > c_{\alpha/2}$, that is, if $Z^* > c_{\alpha/2}$ or if $Z^* < -c_{\alpha/2}$. Therefore, the power of the one-tailed test that $\mu = \mu_0$ is given by

$$p = 1 - \Phi(c_{\alpha/2} - \lambda^*) + \Phi(-c_{\alpha/2} - \lambda^*). \tag{24}$$

Conventions for heterogeneity. Equations 23 and 24 provide expressions for the power of the random-effects test, but they are not useful unless a value of λ^* can be computed, which in turn depends on both the conditional variances $v_1, \ldots, v_k$; the mean $\mu_{\cdot}$; and the between-studies variance component τ^2. As in the homogeneity test in the fixed-effects case, we adopt the convention with a common value of v that $\tau^2 = .33v = v/3$ is a small degree of heterogeneity, $\tau^2 = .67v = 2v/3$ is a medium degree of heterogeneity, and $\tau^2 = v$ is a large degree of heterogeneity. We can take all the v_i values as approximately equal, which gives $v_{\cdot}^*$ approximately equal to $(v + \tau^2)/k$, where v is the common value of the v_i values. Note that if the v_i values are not identical, $(v + \tau^2)/k$ will be strictly larger than $v_{\cdot}^*$, and using $(v + \tau^2)/k$ in place of $v_{\cdot}^*$ in power calculations will result in an underestimate of the statistical power.

Example: Standardized mean differences. Suppose we wanted to estimate a random-effects model with the studies of the effects of childhood sexual abuse, using information from Neumann et al. (1996). The power for the test that the overall mean effect is zero requires knowledge of $v_\cdot^*$ which depends on τ^2, the variance component; v, the common variance of the study effect sizes; and k, the number of studies. In our earlier example, we estimated the mean effect size as 0.20, the minimum value of the effect size that is of substantive interest. We also estimated the common variance of effect sizes across studies, $v_i = 0.111$, a value based on a within-study sample where $N = 48$, $n_{\mathrm{P}} = 12$, and $n_{\mathrm{N}} = 36$. To estimate τ^2, we need to know Q as well as the values of the individual weights for each study (the variance of the effect sizes for each study). We have one of two options here: (a) Estimate τ^2 from Equation 16 by assuming equal weights for each study based on v, or (b) use the convention for small, medium, and large degrees of heterogeneity with τ^2 equal to $v/3$, $2v/3$, and v, respectively. Using (b), we find that a small degree of heterogeneity represents a value of $\tau^2 = .33v = .33 * 0.111 = 0.037$, a medium degree of heterogeneity a value of $\tau^2 = .67v = .67 * 0.111 = 0.074$, and a large degree of heterogeneity a value of $\tau^2 = v = 0.111$. We can then estimate $v_\cdot^* = (v + \tau^2)/k$ for our three values of τ^2, or $v_\cdot^* = (0.111 + 0.037)/18 = 0.0082$ for small heterogeneity, $v_\cdot^* = (0.111 + 0.074)/18 = 0.0103$ for medium heterogeneity, and $v_\cdot^* = (0.111 + 0.111)/18 = 0.012$ for large heterogeneity.

When the heterogeneity is small, the mean value of the test statistic, Z, is given by $\lambda = (0.20 - 0.0)/\sqrt{0.0082} = 2.21$. The power of the one-tailed test at significance level $\alpha = .05$ given in Equation 23 is $p = 1 - \Phi(1.64 - 2.21) = 1 - \Phi(-0.57) = 1 - 0.28 = .72$. With a medium degree of heterogeneity, the mean value of the test statistic, Z, is given by $\lambda = (0.20 - 0.0)/\sqrt{0.0103} = 1.97$. The power at significance level $\alpha = .05$ given in Equation 23 is $p = 1 - \Phi(1.64 - 1.97) = 1 - \Phi(-0.33) = 1 - 0.37 = .63$. A large degree of heterogeneity gives a value of $\lambda = (0.20 - 0.0)/\sqrt{0.012} = 1.82$, with a power of $p = 1 - \Phi(1.64 - 1.82) = 1 - \Phi(-0.18) = 1 - 0.43 = .57$. With approximately 18 studies, a mean effect size of 0.20, and within-study sample size of 48, we have power ranging from .57 to .72 for detecting a mean effect size different from zero in a random-effects model. Note that with a larger degree of heterogeneity, we have less power to detect a nonzero mean in random effects than in fixed effects.

The power of the two-tailed test given in Equation 24 with a small amount of heterogeneity is $p = 1 - \Phi(1.96 - 2.21) + \Phi(-1.96 - 2.21) = 1 - \Phi(-0.25) + \Phi(-4.17) = 1 - 0.40 - 0.00 = .60$. A medium amount of heterogeneity gives the power of the two-tailed test as $p = 1 - \Phi(1.96 - 1.97) + \Phi(-1.96 - 1.97) = 1 - \Phi(-0.01) + \Phi(-3.93) = 1 - 0.50 - 0.00 = .50$. For a two-tailed test with a large amount of heterogeneity, we have power of $p = 1 - \Phi(1.96 - 1.82) + \Phi(-1.96 - 1.82) = 1 - \Phi(0.14) + \Phi(-3.78) = 1 - 0.56 + 0.00 = .44$. With a two-tailed test, we have less power than in the one-tailed test for detecting a difference of 0.20 between exposed and nonexposed adults. Note that even with a small amount of heterogeneity, we have little power in the random-effects case with a nondirectional hypothesis.

Testing the Significance of the Effect Size Variance Component

The test that $\tau^2 = 0$ in the random-effects model is the same as the test of homogeneity in the fixed-effects model using the Q statistic. The reason is that if $\tau^2 = 0$, then

$$\theta_1 = \theta_2 = \ldots = \theta_k = \mu;$$

thus the effect size parameters are fixed, but unknown, constants. This is analogous to the situation with F tests in one-way random- and fixed-effects ANOVAs. In the ANOVA, the null distributions of the test statistics are identical, but the nonnull distributions of the F ratios differ. Similarly, although the null distributions of the Q statistics are identical in fixed and random effect size models, the nonnull distributions of Q differ under the two models.

The test statistic Q has the chi-square distribution with $k - 1$ degrees of freedom when the null hypothesis that $\tau^2 = 0$ is true. When the null hypothesis of homogeneity is not true, that is, when $\tau^2 > 0$, then Q has a distribution of rather complex form (a weighted linear combination of chi-square distributions) that is not extensively tabulated. However, an approximation to that distribution that is adequate for estimating statistical power is known.

In the case that the conditional variances are equal, that is, $v_1 = \ldots v_k = v$, then Q has a distribution that is $(v + \tau^2)/v = 1 + \tau^2/v$ times a central chi-square, with $(k - 1)$ degrees of freedom, so the power of the test that $\tau^2 = 0$ is

$$1 - F[c_\alpha v/(v + \tau^2)|k - 1; 0], \tag{25}$$

where $F(x|\upsilon; 0)$ is the cumulative distribution function of a central chi-square with υ degrees of freedom and c_α is the $100(1 - \alpha)$ percentile point of the chi-square distribution with $k - 1$ degrees of freedom.

When the conditional variances are unequal, an approximation to the distribution of Q can be derived using a method (Satterthwaite, 1946) that approximates the distribution of Q by a gamma random variable with mean and variance equal to that of Q. Because $\hat{\tau}^2$ is an unbiased estimator of τ^2 (except for truncation), the mean μ_Q of Q under this model is

$$\mu_Q = c\tau^2 + (k - 1),$$

where c is given by Equation 17 and the variance of σ_Q^2 of Q is given as (Hedges & Vevea, 1998)

$$\sigma_Q^2 = 2(k-1) + 4\left(\sum w_i - \frac{\sum w_i^2}{\sum w_i}\right)\tau^2 + 2\left[\sum w_i^2 - \frac{\sum w_i^3}{\sum w_i} + \frac{\left(\sum w_i^2\right)^2}{\left(\sum w_i\right)^2}\right]\tau_4. \tag{26}$$

Then the distribution of Q is approximated as a gamma with shape parameter r given by

$$r = \mu_Q^2/\sigma_Q^2$$

and scale parameter m given by

$$m = \mu_Q/\sigma_Q^2.$$

The power of the test that $\tau^2 = 0$ in the random-effects case is

$$1 - F(c_\alpha|r, m), \tag{27}$$

where $F(x|r, m)$ is the cumulative distribution function of a gamma variate with shape parameter r and scale parameter m and c_α is the $100(1 - \alpha)$ percentile point of the chi-square distribution with $k - 1$ degrees of freedom.

For an example, we can return to the Kendall-Tackett et al. (1993) qualitative review, where we assumed that we had $k = 10$ studies with a common sample size of 25, giving us a common value of v (for the transformed Fisher's z scores) as $v = 1/(25 - 3) = 1/22 = 0.045$. For a small amount of variability, we assume the variance component $\tau^2 = 0.33\ (v) = 0.33\ (0.045) = 0.015$. Because we are concerned only with the estimate of τ^2, we do not need to posit an estimate of z, the mean effect size. The power of the test for $\alpha = .05$ for small heterogeneity is given in Equation 25 as

$$\begin{aligned} p &= 1 - F[c_\alpha v/(v + \tau^2)|k - 1; 0] \\ &= 1 - F[(16.92 * 0.045)/(0.045 + 0.015)|9; 0] \\ &= 1 - F(12.69|9; 0) \\ &= 1 - 0.82 \\ &= .18, \end{aligned}$$

where 16.92 is the $100(1 - \alpha)$ percentile point of the chi-square distribution with $10 - 1 = 9$ degrees of freedom. A medium amount of heterogeneity increases the power of the test to $p = 1 - F(10.13|9; 0) = 1 - 0.66 = .34$. When the studies have a large amount of heterogeneity, the power of the test is given as $p = 1 - F(8.46|9; 0) = 1 - 0.51 = .49$. The greater the differences between the studies' effect sizes, the more likely we are to detect a significant value for the variance component though none of these values are high. In general, the power as estimated from both the Kendall-Tackett et al. (1993) and Neumann et al. (1996) reviews is not large for tests of the random-effects model.

Contrasts Among Effect Sizes

Omnibus tests for homogeneity can reveal that the effect parameters are not all the same, but they are not useful for revealing the specific pattern of mean differences that might be present. For example, the Q statistic might reveal that there was variation in the effects, but the omnibus statistic gives no insight about which studies might be associated with the largest effect sizes. In other cases, the omnibus test statistic may not be significant but we may wish to test for a specific difference that the omnibus test may not have been powerful enough to detect. As in conventional ANOVA, contrasts or comparisons are used to explore the differences among group means (Hedges & Olkin, 1985; Rosenthal & Rubin, 1982). Contrasts can be used in precisely the same way to examine patterns among group mean effect sizes in meta-analysis. In fact all of the strategies used for selecting contrasts in ANOVA (such as orthogonal polynomials to estimate trends, Helmert contrasts to discover discrepant groups, etc.) are also applicable in meta-analysis.

A contrast (parameter) is just a linear combination of group means

$$\gamma = c_1\bar{\theta}_1 + \ldots + c_k\bar{\theta}_k, \tag{28}$$

where the coefficients $c_1, \ldots, c_k$ (called the contrast coefficients) are known constants that satisfy the constraint $c_1 + \ldots + c_k = 0$ and are chosen so that the value of the contrast will reflect a particular comparison or pattern of interest. For example, the coefficients $c_1 = 1, c_2 = -1, c_3 = \ldots = c_k = 0$ might be chosen so that the value of the contrast is the difference between the effect size $\overline{\theta}_1$ of Study 1 and the effect size $\overline{\theta}_2$ of Study 2. Sometimes we refer to a contrast among population effect sizes as a *population contrast* or a *contrast parameter* to emphasize that it is a function of population parameters and to distinguish it from *estimates* of the contrast. The contrast parameter specified by coefficients $c_1, \ldots, c_k$ is usually estimated by a sample contrast

$$G = c_1\overline{T}_1 + \ldots + c_k\overline{T}_k. \quad (29)$$

The estimated contrast G has a normal sampling distribution with variance v_G given by

$$v_G = c_1^2 v_1 + \ldots + c_k^2 v_k. \quad (30)$$

Note that although the notation used here for contrasts suggests that they compare individual study effect sizes, they can be used to compare groups of studies or to compare a single study with a group mean. All that is required is the appropriate selection of contrast coefficients to define the "groups" of studies involved.

Because the estimated contrast G has a normal distribution with known variance v_G, tests of statistical significance are relatively easy to construct. Note, however, that just as with contrasts in ordinary ANOVA, test procedures differ depending on whether the contrasts were planned or were selected using information from the data. Here we discuss only procedures for testing planned comparisons.

A test of the null hypothesis that $\gamma = 0$ uses the statistic

$$Z_G = G/\sqrt{v_G}. \quad (31)$$

A one-tailed test of the null hypothesis that $\gamma = 0$ uses the statistic Z_G given above but rejects the hypothesis that $\gamma = 0$ and declares the contrast to be significant at level of significance alpha if $Z_G > c_\alpha$. The two-tailed test rejects the null hypothesis that $\gamma = 0$ (declares the contrast to be significant at the level of significance alpha) if $|Z_G| > c_{\alpha/2}$, where $c_{\alpha/2}$ is the 100 $(1 - \alpha/2)$ percent point of the standard normal distribution.

When the null hypothesis that $\gamma = 0$ is true, Z_G has the standard normal distribution, but when the null hypothesis is false, Z_G has mean $\gamma/\sqrt{v_G}$ and variance 1. Because the one-tailed test at significance level alpha rejects the null hypothesis if $Z_G > c_\alpha$, the power of the one-tailed test that $\gamma = 0$ is given by

$$p = 1 - \Phi(c_\alpha - \gamma/\sqrt{v_G}), \quad (32)$$

where $\Phi(x)$ is the standard normal cumulative distribution function.

The computation of power in the two-tailed test is only slightly more complicated. Because the two-tailed test at significance level alpha rejects the null hypothesis if $|Z_G| > c_{\alpha/2}$, that is, if $Z_G > c_{\alpha/2}$ or if $Z_G < -c_{\alpha/2}$, the power of the two-tailed test that $\gamma = 0$ is given by

$$p = 1 - \Phi(c_{\alpha/2} - \gamma/\sqrt{v_G}) + \Phi(-c_{\alpha/2} - \gamma/\sqrt{v_G}). \quad (33)$$

For example, the Neumann et al. (1996) review of the effects of childhood sexual abuse in women reported on a number of study moderators and their relationship to effect size. One important moderator is the locus of abuse. As Rind et al. (1998) argued, studies with participants restricted to those abused by close family members may estimate larger effect sizes than studies in which participants include those abused by strangers or by family members. We can use the results of Neumann et al. to get an idea of whether we are likely to have enough power to detect this difference.

As can be seen in Table 1, Neumann et al. (1996) identified 8 studies in which the locus of abuse was intrafamilial, with an average effect size of 0.47 and a 95% confidence interval ranging from 0.29 to 0.69. Another 29 studies included participants reporting a mixed locus of abuse, with an average effect size of 0.35 and a 95% confidence interval ranging from 0.30 to 0.39. We could posit a G, the sample contrast value, of 0.20, as the minimum difference we care about detecting in the meta-analysis. We need to compute the common variance for the effect sizes in each group of studies. For the intrafamilial abuse studies, we have the lower bound of the 95% confidence interval equal to $0.29 = \delta_L = d. - c_{\alpha/2}\sqrt{v.}$, where $d. = 0.47$ and $c_{\alpha/2} = 1.96$. From this equation, we obtain $v. = [(0.29 - 0.47)/(-1.96)]^2 = (0.0092)^2 = 0.0085$ as our value for the variance of the mean effect size for studies with intrafamilial abuse. For the mixed abuse studies, we have a lower bound of the 95% confidence interval equal to $0.30 = 0.35 - 1.96\sqrt{v.}$, giving $v. = 0.00065$. The variance of the contrast G is $v_G = (1)^2\, 0.0085 + (-1)^2\, 0.00065 = 0.00905$. The power of the one-tailed test that $\gamma = 0$ is given by Equation 32, or

$$\begin{aligned} p &= 1 - \Phi(c_\alpha - \gamma/\sqrt{v_G}) \\ &= 1 - \Phi(1.64 - 0.20/\sqrt{0.00905}) \\ &= 1 - \Phi(1.64 - 2.10) \\ &= 1 - \Phi(-0.46) \\ &= 1 - 0.32 \\ &= .68. \end{aligned}$$

If we found a difference of 0.20, with studies similar to those found by Neumann et al., we would have power of .68 to detect a significant difference between the effect sizes of studies focusing on intrafamilial abuse versus those with mixed abuse. The power of the two-tailed test, if we did not have an idea of the direction of the difference, would give $1 - \Phi(1.96 - 2.10) + \Phi(-1.96 - 2.10) = 1 - \Phi(-0.14) - \Phi(-4.06) = 1 - 0.44 + 0.0 = .36$. We would have a small amount of power to detect a difference of 0.20 with the numbers of studies given by Neumann et al.

Conclusion

Analysis of the power of statistical tests is an important part of planning any scientific research study, including meta-analyses. Computation of statistical power is essential to know whether tests of hypotheses are likely to detect the effects expected (if they obtain in the population). Although many people believe that meta-analyses necessarily have high statistical power, the examples used in this article, which come from published meta-analyses, demonstrate that this is not necessarily the case. Tests of heterogeneity are particularly vulnerable to low statistical power given the degree of heterogeneity that is plausible in most meta-analytic situations.

The inclusion in a meta-analysis of studies with very small sample sizes may have a paradoxical effect of decreasing the power of random-effects tests of the mean effect size. That is, it is possible that small studies may introduce enough heterogeneity into the analysis to more than compensate for the added information about the mean that they provide. Consequently, one may actually achieve higher statistical power by excluding such studies. Although exclusion of small studies cannot increase statistical power in fixed-effects analyses, there are still reasons to consider this option.

If only studies that individually have high statistical power are included in the meta-analysis, the meta-analysis will necessarily (in the case of fixed-effects analyses) or probably (in the case of random-effects analyses without too much heterogeneity) have high power. Power considerations aside, publication selection (the tendency of studies with statistically significant results to be more likely to be published) produces larger bias in studies with low statistical power than studies with high statistical power. Therefore, exclusion of studies with low individual statistical power may provide some protection against the effects of publication selection (Kraemer, Gardner, Brooks, & Yessavage, 1998).

Power computations in primary analysis require knowing at least the effect size and the sample size. In meta-analysis the sample size has two components: the number of studies and the within-study sample size. When the statistical analysis explicitly takes into account the heterogeneity of effects, then information about the heterogeneity of effects is also required in both primary analysis and meta-analysis.

Power analyses for future meta-analyses could be facilitated by more complete reporting of research reviews that might be used to obtain information for power analyses. Perhaps the best information would be a table reporting key substantive characteristics, sample sizes, and (for meta-analyses) effect sizes for all studies in the review. In the absence of such a table, reports of the distribution of sample sizes and the between-studies variance component (for meta-analyses) would be useful. We advocate computing the variance component (not just the standard deviation of the effect size estimates) as a descriptive statistic to describe heterogeneity of effects even if a fixed-effects analysis is used. Reporting these variance components would provide important information for computing power in future random-effects analyses and (across reviews) would increase the available knowledge about typical levels of effect size heterogeneity.

In order for power analysis to be a scientific enterprise, determination of the parameters on which power depends must be taken seriously. The use of data from previous reviews or data from representative samples of studies considered for inclusion in the review corresponds to sound practice in power analysis in primary research. Informed professional judgment may also be used, but it is difficult to evaluate its accuracy a priori. Simply taking a wild guess and inserting values into the formulas in this article is little better than simply guessing the power.

References

Bornstein, M. (2000). *Power and Precision* (*Version 2.0*) [Computer software]. Englewood, NJ: Biostat.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

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

Diggle, P. J., Liang, K. L., & Zeger, S. L. (1994). *Analysis of longitudinal data.* Oxford, England: Oxford University Press.

Elashoff, J. D. (2000). *nQuery Advisor* (*Version 4.0*) [Computer software]. Saugus, MA: Statistical Solutions.

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

Hedges, L. V. (1982b). Fitting categorical models to effect size from a series of experiments. *Journal of Educational Statistics, 7,* 119–137.

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

Hedges, L. V., & Olkin, I. (1985). *Statistical models for meta-analysis.* New York: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random effects models in meta-analysis. *Psychological Methods, 3,* 486–501.

Kendall-Tackett, K. A., Williams, L. M., & Finkelhor, D. (1993). Impact of sexual abuse on children: A review and synthesis of recent empirical studies. *Psychological Bulletin, 113,* 164–180.

Kraemer, H. C., Gardner, C., Brooks, J. O. I., & Yessavage, J. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data.* Pacific Grove, CA: Brooks/Cole.

Neumann, D. A., Houskamp, B. M., Pollock, V. E., & Briere, J. (1996). The long-term sequelae of childhood sexual abuse in women. *Child Maltreatment, 1,* 6–16.

O'Brien, R. G. (1998). *UnifyPow* [Computer software]. Cleveland, OH: Author.

Patnaik, P. B. (1949). The non-central chi-square and *F*-distributions and their applications. *Biometrika, 36,* 202–232.

Rind, B., Tromovitch, P., & Bauserman, R. (1998). A meta-analytic examination of assumed properties of child sexual abuse using college samples. *Psychological Bulletin, 124,* 22–53.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500–504.

SAS Institute. (1990). *SAS language: Reference* [Software manual]. Cary, NC: Author.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114.

Schmidt, F. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

Schmidt, F., & Hunter, J. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 65,* 643–661.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes from two-level research. *Journal of Educational Statistics, 18,* 237–259.

SPSS. (1999). *SPSS Base 10.0 applications guide* [Software manual]. Chicago: Author.

Received September 21, 1999
Revision received December 31, 2000
Accepted April 4, 2001 ■