Psychological Methods
2003, Vol. 8, No. 3, 243–253

1082-989X/03/$12.00 DOI: 10.1037/1082-989X.8.3.243

How Meta-Analysis Increases Statistical Power

Lawrence D. Cohn
University of Texas at El Paso

Betsy J. Becker
Michigan State University

One of the most frequently cited reasons for conducting a meta-analysis is the increase in statistical power that it affords a reviewer. This article demonstrates that fixed-effects meta-analysis increases statistical power by reducing the standard error of the weighted average effect size ($\overline{T}.$) and, in so doing, shrinks the confidence interval around $\overline{T}.$. Small confidence intervals make it more likely for reviewers to detect nonzero population effects, thereby increasing statistical power. Smaller confidence intervals also represent increased precision of the estimated population effect size. Computational examples are provided for 3 effect-size indices: *d* (standardized mean difference), Pearson's *r,* and odds ratios. Random-effects meta-analyses also may show increased statistical power and a smaller standard error of the weighted average effect size. However, the authors demonstrate that increasing the number of studies in a random-effects meta-analysis does not always increase statistical power.

Since 1976, thousands of meta-analyses have been published in a variety of disciplines, including medicine, psychology, environmental science, and education. One of the most frequently cited reasons for conducting a meta-analysis is the increase in statistical power that it affords a reviewer. The following passages illustrate this point:

> The problems created by low statistical power in individual studies are central to the need for meta-analysis (Hunter & Schmidt, 1990, p. 75);

> Quantitative reviews or meta-analysis have been likened to a tower of statistical power that allows researchers to rise above the body of evidence (Murlow, 1995, p. 4);

> Compared with statistical analyses in primary studies, statistical power will typically be much higher in meta-analyses (Matt & Cook, 1994, p. 510);

> Among the many virtues that have been extolled for meta-analysis, the main appeal is that it can convert existing things into something better. "Significance" can be attained statistically when small group sizes are pooled into big ones (Feinstein, 1995, p. 71).

Illustrative of these claims is a meta-analysis of 33 clinical trials examining the effect of administering intravenous streptokinase to patients who were hospitalized for acute myocardial infarction (Lau et al., 1992). Twenty-seven of the 33 trials failed to identify significant treatment effects on patient mortality, yet a meta-analysis of these same studies revealed that streptokinase significantly reduces the odds of dying by approximately 20%.

How does meta-analysis increase statistical power and thereby help reviewers detect population effects or relationships in a set of studies when those effects remain undetected within the individual studies themselves? Some investigators erroneously assume that the increased statistical power results from a simple pooling of subjects across studies, resulting in a single large sample and increased statistical power. Meta-analysis, however, does not pool subjects into a single sample. Indeed the latter strategy is not recommended (Bravata & Olkin, 2002) and can result in outcomes characterized by Simpson's paradox (Simpson, 1951).

Editor's Note. William R. Shadish served as action editor for this article.—SGW

Lawrence D. Cohn, Department of Psychology, University of Texas at El Paso; Betsy J. Becker, Department of Counseling, Educational Psychology, and Special Education, Michigan State University.

Preparation of this article was partially supported by a Faculty Development Leave awarded to Lawrence D. Cohn by the University of Texas at El Paso and a concurrent visiting faculty appointment in Developmental and Educational Psychology, Department of Psychology, Leiden University, the Netherlands.

Correspondence concerning this article should be addressed to Lawrence D. Cohn, Department of Psychology, University of Texas at El Paso, El Paso, Texas 79968-0553. E-mail: Lcohn@utep.edu

Other researchers hold more skeptical views, maintaining that meta-analysis is a form of statistical alchemy, converting nonsignificant findings into significant ones (Feinstein, 1995).

Although several articles have addressed issues related to meta-analysis and statistical power (e.g., Hedges & Pigott, 2001; Hunter & Schmidt, 1990; Strube, 1985; Strube & Miller, 1986), no article has explained how meta-analysis increases the statistical power of tests of overall treatment effects and relationships. This article addresses this gap in explanation. We begin with a brief discussion of power in general and in the meta-analytic context, then we illustrate how meta-analysis increases power in three syntheses using different effect-size indices.

Statistical Power

Statistical power refers to the likelihood of detecting, within a sample, an effect or relationship that exists within the population. More formally stated, "the power of a statistical test of a null hypothesis is the probability that it will lead to the rejection of the null hypothesis, i.e., the probability that it will result in the conclusion that the phenomenon exists" (Cohen, 1988, p. 4). Early work by Cohen (1962) revealed that behavioral science research was plagued by low statistical power. For example, he reported that the median level of statistical power associated with studies published in the 1960 volume of the *Journal of Social and Abnormal Psychology* was only .46 (assuming a two-tailed test, $\alpha = .05$, and a medium effect size); more recent assessments reveal similarly low estimates of statistical power (Cohen, 1992; Sedlmeier & Gigerenzer, 1989). Traditional narrative reviews are unable to distinguish between null findings that result from low power and null findings that reflect a genuine absence of population effects, sometimes leading reviewers to erroneously conclude that a body of evidence does not support a proposed relationship or treatment effect. Indeed, if the statistical power of each study in a review is less than .50, then simply tabulating the number of significant findings (i.e., vote counting) leads "to the wrong decision more often as the amount of evidence (number of studies) increases" (Hedges & Olkin, 1985, p. 50; see also Hedges & Olkin, 1980; Hunter & Schmidt, 1990). A meta-analysis of such studies, however, can increase the likelihood of detecting population effects. The following discussion explains why this occurs.

Meta-Analysis

One typical goal of meta-analysis is to estimate a population effect-size parameter, θ. In this article we are concerned with the power of tests of null hypotheses about θ. A second goal of meta-analysis is to increase the precision of our estimate of θ. Two types of analyses can be used to reach these goals: fixed-effects analyses and random-effects analyses. The decision to use a random- or fixed-effects meta-analysis depends on a number of issues. For illustrative purposes, we subject each of three data sets to a fixed- or random-effects analysis without considering such issues as heterogeneity, publication bias, or other theoretical concerns that would typically influence a reviewer's choice of models. We begin with a discussion of power under the fixed-effects model and then turn to a discussion of random-effects analyses.

Fixed-Effects Models

The simplest fixed-effects analysis assumes that all samples arise from one population with a common parameter θ. Equivalently we may say that we are assuming that a fixed-effects model applies in such cases. Each study in the analysis contributes to the estimate of θ by providing a sample effect-size index, T_i (e.g., a standardized mean difference or correlation). Effect sizes with small variances provide better (more precise) estimates of the population effect size (θ) than do effects with large variances. Thus, effect sizes with small variances are weighted more heavily when estimating θ. Notably, the variance (v_i) of a sample effect size is always a function of the sample size of the study from which it was calculated. (We show formulas for several instantiations of v_i below.)

When the fixed-effects model applies, the weighted average effect size ($\overline{T}.$) provides the best estimate of the population effect size θ (Hedges & Olkin, 1985). The weighted average effect size is defined (e.g., by Hedges & Olkin, 1985) as

$$\overline{T}. = \sum w_i T_i / \sum w_i,$$

where T_i is the effect size computed from the ith study, and w_i is the weight assigned to the effect size in the ith study. For the fixed-effects model the weight w_i is, in general, defined as $w_i = 1/v_i$. Each sample effect size (T_i) contributing to the estimate of $\overline{T}.$ is weighted by the inverse (i.e., reciprocal) of its variance. Weighting by the inverse of the variance ensures that effect sizes with small variances contribute

more to the weighted average effect size than do effect sizes with large variances.

The variance ($V_.$) of the weighted average effect size is approximated by

$$V_. = 1/\sum w_i,$$

where w_i represents the weight assigned to the ith effect size (Hedges & Olkin, 1985). Under the fixed-effects model the 95% confidence interval (CI) constructed around $\overline{T}_.$ is defined as

$$\overline{T}_. \pm 1.96 \sqrt{V_.}.$$

Notably, under the fixed-effects model, the value of $V_.$ becomes smaller as the total number of subjects included in a meta-analysis increases (all other things being equal). Hence the standard error ($\sqrt{V_.}$) of the weighted average effect size also becomes smaller as the number of subjects in a meta-analysis increases. As the standard error ($\sqrt{V_.}$) of the weighted average effect size becomes smaller, the confidence interval around $\overline{T}_.$ (the weighted average effect size) must also become smaller. Small confidence intervals are less likely to include zero (assuming θ does not equal zero), making it easier to detect nonzero population effects. Small confidence intervals also correspond with more precise estimates of θ. Indeed, it is the smaller confidence intervals that many investigators regard as the main advantage associated with meta-analysis.

The test of a nonzero effect (i.e., a test of H_0: $\theta = 0$) can be computed as $Z = \overline{T}_./\sqrt{V_.}$, which under certain standard assumptions has a standard normal distribution under the null hypothesis. Hedges and Pigott (2001) showed how to compute power levels for a variety of tests in meta-analysis, including this test of the population effect. They noted that if the null hypothesis is false, and θ takes on some other value (say θ_1), then the mean of the distribution of the Z test is not zero but rather is

$$\lambda = (\theta_1 - \theta)/\sqrt{V_.}.$$

Hedges and Pigott showed that the power of a two-sided Z test to detect an effect of size θ_1 is

$$\text{Power} = 1 - \Phi(C_{\alpha/2} - \lambda) + \Phi(-C_{\alpha/2} - \lambda),$$

where $\Phi(x)$ is the standard normal cumulative distribution function, and $C_{\alpha/2}$ is the standard normal critical value for the two-sided test at level α (i.e., we compare $[Z]$ to $C_{\alpha/2}$).

According to this complex formula, power increases as λ increases, which occurs either when variability (i.e., $V_.$) decreases or the population effect to be detected (θ_1) deviates from θ. Below we have computed power values for the test of H_0: $\theta = 0$ at a value of θ_1 equal to the mean effect that is estimated from the full meta-analysis used in each example. We chose this value simply for convenience. An argument can be made that, in real practice, one should choose a value of θ_1 on the basis of theoretical considerations, practical import, or available data. In addition, in real practice the meta-analyst would not usually know the value of the mean effect before conducting the synthesis (when power might be calculated).

Fixed-Effects Meta-Analysis and Statistical Power

Standardized Mean Differences

We consider first the example of the standardized-mean-difference effect size, which compares the average performances of two groups on some outcome (here denoted as X). Let g_i be defined as

$$g_i = (\overline{X}_{1i} - \overline{X}_{2i})/S_i,$$

where $\overline{X}_{1i}$ and $\overline{X}_{2i}$ are the means of the two comparison groups, and S_i is the pooled within-groups standard deviation. In general (per Hedges, 1981) we typically unbias the estimator g_i by means of

$$d_i = c(m_i)g_i \approx 1 - 3/(4\,m_i - 1)\,g_i,$$

where m_i is the appropriate degrees of freedom—for the two-sample comparison it is $m_i = n_{1i} + n_{2i} - 2$. So, if the index of effect (T_i above) is an unbiased standardized mean difference, d_i, then the variance of d_i is defined as

$$v_i = [(n_{1i} + n_{2i}) \div (n_{1i} \times n_{2i})] + [d_i^2 \div [2(n_{1i} + n_{2i})]],$$

where n_{1i} and n_{2i} represent the sample sizes of the two groups in the ith study (Hedges & Olkin, 1985). As the sample size increases, v_i must necessarily decrease in value. Also w_i is defined as

$$w_i = 1/v_i = \{[(n_{1i} + n_{2i}) \div (n_{1i} \times n_{2i})] + [d_i^2 \div [2(n_{1i} + n_{2i})]]\}^{-1}.$$

The weighted average effect size ($\bar{d}_.$) is itself a sample statistic with a known variance and a known sampling distribution (Hedges & Becker, 1986). The weighted average effect size for standardized mean differences is defined as

$$\bar{d}_. = \sum w_i d_i / \sum w_i,$$

where d_i is the standardized mean difference computed from the ith study, and w_i given above is the weight assigned to the standardized mean difference in the ith study. For large samples ($n_{ji} > 10$ per group in each study), $\bar{d}.$ is approximately normally distributed (Hedges & Olkin, 1985). The variance ($V.$) of the weighted average effect size is approximated by $V. = 1/\sum w_i$. The 95% CI constructed around $\bar{d}.$ is defined as above: $\bar{d}. \pm 1.96 \sqrt{V.}$.

The data presented in Table 1 illustrate the relationship between meta-analysis and statistical power for standardized-mean-difference data. The data are adapted from a meta-analysis of gender differences in risk taking (Byrnes, Miller, & Schafer, 1999). Table 1 summarizes the subset of 12 findings regarding sex differences in self-reported risky driving behaviors among 14- to 21-year-old adolescents. To meet the assumption of independence, we used only one effect size per sample in the present analysis. We subjected these data to a cumulative meta-analysis (Lau et al., 1992) to illustrate how meta-analysis increases statistical power. We have ordered and accumulated the studies, according to publication date, to show how power increases as studies accumulate over time.

If the meta-analysis only included the first study (Flaherty and Arenson, 1978; as cited in Byrnes et al., 1999), then the 95% CI would equal $\bar{d}. \pm 0.483$ and the power to detect an effect equal to $\delta = 0.36$ standard deviations would be 0.31 (see Table 1, last two columns). If the first two studies (Flaherty & Arenson, 1978, and Moore and Rosenthal, 1992; as cited in Byrnes et al., 1999) are included in the meta-analysis then the 95% CI is $\bar{d}. \pm 0.254$, and the power increases to just under 80%. Similarly, if the first three studies are included in the meta-analysis then the 95% CI is $\bar{d}. \pm 0.101$, and if four studies are included in the analysis then the 95% CI is $\bar{d}. \pm 0.073$, and the power has risen to be essentially 1.0. We would be almost sure to detect a population effect of 0.36 standard deviations even with only the first 3 of the 12 studies that have been conducted. If all 12 studies are included in the meta-analysis then the 95% CI is $\bar{d}. \pm 0.046$. Note that power increases as each new sample is added to the meta-analysis (see the next-to-last column in Table 1). This is reflected in the confidence intervals that become progressively smaller, as is readily apparent in Figure 1A. Note, also, that as studies are added, the combined estimate of the size of the gender difference stabilizes and converges to an estimate near the mean of the 12 effects; this is also evident in the figure. As smaller confidence intervals converge on a population effect that is nonzero, the likelihood of detecting the nonzero effect increases; that is, small confidence intervals increase statistical power when the true effect θ is nonzero. A small confidence interval also increases the precision of our estimate, narrowing the range of potential population values.

Pearson Correlations

Next we consider the case where our T_i is a Pearson product–moment correlation, that is, $T_i = r_i$. Under

Table 1
A Cumulative Meta-Analysis of Sample d_is Assessing Gender Differences in Risky Driving Behaviors (Fixed-Effects Model)

Study date	Standardized mean difference d_i	n_M	n_F	Cumulative n	Cumulative weighted average effect size $\bar{d}.$	95% CI	Power to detect $\delta = 0.36$
1978	0.03	28	40	68	0.03	$\bar{d}. \pm 0.483$	.31
1992	0.53	71	118	257	0.39	$\bar{d}. \pm 0.254$	.79
1992	0.67	678	673	1,608	0.63	$\bar{d}. \pm 0.101$	>.99
1993	0.06	704	672	2,984	0.36	$\bar{d}. \pm 0.073$	>.99
1994	0.33	108	125	3,217	0.35	$\bar{d}. \pm 0.070$	>.99
1995	–0.08	139	147	3,503	0.32	$\bar{d}. \pm 0.067$	>.99
1995	0.75	55	101	3,659	0.33	$\bar{d}. \pm 0.066$	>.99
1995	0.48	303	376	4,338	0.36	$\bar{d}. \pm 0.061$	>.99
1996	0.61	88	178	4,604	0.37	$\bar{d}. \pm 0.059$	>.99
1996	0.22	559	373	5,536	0.34	$\bar{d}. \pm 0.054$	>.99
1996	0.23	156	147	5,839	0.34	$\bar{d}. \pm 0.052$	>.99
1996	0.43	906	880	7,625	0.36	$\bar{d}. \pm 0.046$	>.99

Note. Sample sizes and effect sizes in columns 2, 3, and 4 are as cited in "Gender Differences in Risk Taking: A Meta-Analysis," by J. P. Byrnes, D. C. Miller, & W. D. Schafer, 1999, *Psychological Bulletin, 125,* Table 1, pp. 374–376. Copyright 1999 by the American Psychological Association. Adapted with permission of author. Positive d_i values indicate more risky behaviors for males. n_M = male sample size; n_F = female sample size; $\bar{d}.$ = weighted average effect size; CI = confidence interval.

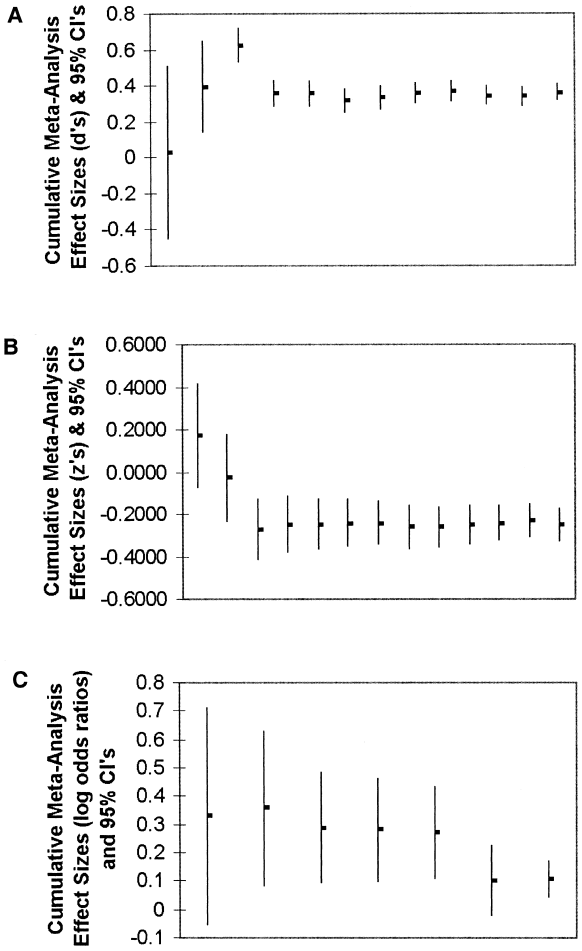

Figure 1. Cumulative meta-analyses and 95% confidence intervals (CIs).

the fixed-effects model, each sample correlation that is included in a meta-analysis provides an estimate of the population correlation ρ. When $\rho \neq 0$ the sampling distribution of r departs from normality, making it difficult to construct confidence intervals around the estimate of ρ. In meta-analysis, therefore, sample correlations are often replaced by Fisher's z-transformed correlations (z_i), which have a known variance and a sampling distribution that approximates a normal distribution for sample sizes of at least 20 (Hedges & Olkin, 1985). Consider the case where the correlation r_i in study i is based on n_i pairs of scores. The variance (v_i) of Fisher's transformation z_i equals $1/(n_i - 3)$, the associated weight w_i equals $n_i - 3$ (i.e., the inverse of the variance), and the weighted average of the z-transformed correlations is defined as

$$\bar{z}_{r.} = \sum w_i z_i \div \sum w_i.$$

The weighted average of the z-transformed correlations ($\bar{z}_{r.}$) is itself a sample statistic with a known variance and sampling distribution that is approximately normally distributed (Hedges & Olkin, 1985). The variance ($V.$) of the weighted average z-transformed correlation is approximated by

$$V. = 1/\sum w_i = 1/\sum (n_i - 3) = 1/(N - 3k),$$

where w_i is the weight assigned to the ith z value (that is, the z-transformed correlation from study i), N is the total number of subjects in the meta-analysis ($N = \sum n_i$), and k is the number of studies. This representation of $V.$ shows clearly how $V.$ becomes smaller as the number of subjects in the meta-analysis (N) increases. The 95% CI constructed around $\bar{z}_{r.}$ is defined as

$$\bar{z}_{r.} \pm 1.96 \sqrt{V.}.$$

Hence the standard error ($\sqrt{V.}$) of the weighted average of the z-transformed correlations ($\bar{z}_{r.}$) also becomes smaller as the number of subjects in a meta-analysis increases. As the standard error ($\sqrt{V.}$) of the weighted average correlation gets smaller, the confidence interval around $\bar{z}_{r.}$ also gets smaller, making it less likely that zero is included within the CI (when $\rho \neq 0$).

The data presented in Table 2 illustrate how meta-analysis increases the statistical power associated with a set of sample correlations. The data are adapted from a narrative review of studies that investigated the relationship between sensation seeking scores and levels of monoamine oxidase, a brain enzyme (Zuckerman, 1994, p. 298). Six of the 13 correlations reported in the review are nonsignificant (i.e., have individual p values above .05). We subjected the 13 z-transformed correlations to the type of cumulative meta-analysis described in the previous section. Power values are computed for detecting a population z value of $z_\rho = 0.25$ (the mean effect found for all 13 studies). We have presented the means and confidence intervals in the z metric, but for the values we show the transformation back to the correlation scale would not change the values much. For instance, a mean Fisher $\bar{z}_{r.}$ equal to 0.25 corresponds to a mean correlation of .245.

Results are presented in the last columns of Table 2. When the meta-analysis is restricted to Study 1, then the 95% CI is $\bar{z}_{r.} \pm 0.249$ with power of 50%; when the meta-analysis is restricted to Studies 1 and 2 then the 95% CI is $\bar{z}_{r.} \pm 0.208$ with power equal to .65, and when the meta-analysis is restricted to Stud-

Table 2
A Cumulative Meta-Analysis of Sample Correlations Between Sensation Seeking Scores and Monoamine Oxidase

Study date	r_i	n_i	Cumulative n	Cumulative weighted average effect size $\bar{z}_{r.}$	95% CI	Power to detect $z_\rho = -.25$
1977	.17	65	65	.17	$\bar{z}_{r.} \pm 0.249$	.50
1977	−.45**	30	95	−.03	$\bar{z}_{r.} \pm 0.208$	.65
1978	−.47**	93	188	−.27	$\bar{z}_{r.} \pm 0.146$	.92
1983	−.13	36	224	−.25	$\bar{z}_{r.} \pm 0.135$	.95
1987	−.24*	57	281	−.25	$\bar{z}_{r.} \pm 0.120$	.98
1987	−.15	30	311	−.24	$\bar{z}_{r.} \pm 0.114$	.99
1987	−.25	40	351	−.24	$\bar{z}_{r.} \pm 0.108$	>.99
1988	−.66**	13	364	−.26	$\bar{z}_{r.} \pm 0.106$	>.99
1988	−.25	44	408	−.26	$\bar{z}_{r.} \pm 0.100$	>.99
1989	−.23*	58	466	−.25	$\bar{z}_{r.} \pm 0.094$	>.99
1990	−.18*	125	591	−.24	$\bar{z}_{r.} \pm 0.083$	>.99
1990	.18	10	601	−.23	$\bar{z}_{r.} \pm 0.082$	>.99
1990	−.74**	13	614	−.25	$\bar{z}_{r.} \pm 0.082$	>.99

Note. Sample correlations in column 2 are as cited in *Behavioral Expressions and Biosocial Bases of Sensation Seeking* (p. 298), by M. Zuckerman, 1994, New York: Cambridge University Press. Copyright 1994 by Cambridge University Press. Adapted with permission of Cambridge University Press. r_i = Pearson correlation; n_i = sample size; $\bar{z}_{r.}$ = weighted average z-transformed correlation; CI = confidence interval. *$p < .05$. **$p < .01$.

ies 1, 2, and 3 then the 95% CI is $\bar{z}_{r.} \pm 0.146$, and power is .92. Including all 13 studies in the meta-analysis yields the smallest 95% CI: $\bar{z}_{r.} \pm 0.082$ (see Figure 1B). Here, again, the data illustrate how each additional study in a meta-analysis reduces the 95% CI constructed around the weighted average effect size (in this case, the weighted average of the z-transformed correlations). A small confidence interval increases the likelihood of detecting nonzero population correlations, thereby increasing statistical power. A small confidence interval also increases the precision of our estimate, narrowing the range of plausible potential population values. Also again in this data set, the probability of detecting a population effect the size of the mean $\bar{z}_{r.}$ (based on all studies in the meta-analysis) had reached essentially 1.0 when only about half of the studies included in the synthesis had been conducted.

Odds Ratios

Odds ratios are an appropriate effect-size index, T_i, for a meta-analysis of dichotomous data (Fleiss, 1993, 1994; Haddock, Rindskopf, & Shadish, 1998). Each sample odds ratio (or_i) included in a meta-analysis provides an estimate of the population odds ratio (T). The sampling distribution of the odds ratio departs from normality, making it difficult to construct confidence intervals around the estimate of $\bar{T}.$. In a meta-analysis, therefore, each sample odds ratio (or_i) is replaced by the natural log of the odds ratio, which has a known variance and a sampling distribution that approximates a normal distribution (Fleiss, 1994). Let the natural log of the odds ratio be denoted by $T_i = \text{Ln}[or_i]$. The variance (v_i) of $\text{Ln}[or_i]$ is defined as

$$v_i = 1/n_{1i} + 1/n_{2i} + 1/n_{3i} + 1/n_{4i},$$

where n_{1i}, n_{2i}, n_{3i}, and n_{4i} represent the relevant sample sizes in a study (e.g., n_{1i} could represent the number of subjects in a smoking cessation program who stopped smoking, n_{2i} represents the number of participants in the program who did not stop smoking, n_{3i} represents the number of participants in the control group who stopped smoking, and n_{4i} represents the number of participants in the control group who did not stop smoking). The weight (w_i) assigned to a log-transformed odds ratio is defined as $1/v_i$ (i.e., the inverse of the variance). Thus the weighted average effect size (in this case, the weighted average of the natural log-transformed odds ratios) is defined as

$$\bar{L}. = \sum w_i \,\text{Ln}[or_i] \div [\sum w_i].$$

The *weighted average log odds ratio* (i.e., $\bar{L}.$) is also a sample statistic that has a known variance and sampling distribution. $\bar{L}.$ is approximately normally distributed (Fleiss, 1994). The variance ($V.$) of the weighted average of the log odds ratios is approximated by

$$V. = 1/\sum w_i,$$

where w_i represents the weight assigned to the ith sample log-transformed odds ratio. The 95% CI con-

structed around the weighted average of the log odds ratios is

$$\bar{L}. \pm 1.96 \sqrt{V}..$$

Here, too, the value of $V.$ becomes smaller as the number of subjects in the meta-analysis increases. Hence the standard error ($\sqrt{V}.$) of the weighted average of the log odds ratios also becomes smaller as the number of subjects in a meta-analysis increases. As the standard error ($\sqrt{V}.$) of the weighted average-log odds ratio becomes smaller, the confidence interval around $\bar{L}.$ also becomes smaller, reducing the likelihood that zero is included within the CI (when $\text{Ln}[T] \neq 0$). Thus decreases in $V.$ lead to higher power, as was true for the two previous indices.

The data presented in Table 3 illustrate how meta-analysis increases the statistical power associated with a set of sample odds ratios. The data are adapted from a summary of seven studies investigating if mortality rates decline after administering aspirin to patients who suffered myocardial infarctions (Fleiss, 1993; the Appendix provides the relevant summary statistics). Power values are computed for detecting a population log odds ratio of $\text{Ln}[T] = 0.109$ (the mean effect found for all seven studies).

If the cumulative meta-analysis is restricted to Study 1, then the 95% CI is $\bar{L}. \pm 0.386$ with power of about 9%; if the meta-analysis is restricted to Studies 1 and 2 then the 95% CI is $\bar{L}. \pm 0.277$, and power is .12; and if the meta-analysis is restricted to Studies 1, 2, and 3 then the 95% CI is $\bar{L} \pm 0.197$, and power has increased to only 20%. Each additional study reduces still further the 95% CI and increases the power (see Table 3, last two columns). Including all seven studies in the meta-analysis yields the smallest 95% CI: $\bar{L}. \pm 0.065$ (see Figure 1C) and the power has reached about 92%—still not quite as high as for our first two examples. This outcome does not reflect a property of odds ratios; instead, the low power reflects the fact that the effect to be detected here ($\text{Ln}[T] = 0.109$) is relatively small, corresponding to an odds ratio just over one ($T = 1.12$). Small confidence intervals increase the likelihood of detecting odds ratios different from 1.0, thereby increasing statistical power even when the effect to be detected is rather small. Small confidence intervals also increase the precision of our estimate of the population effect size.

Random-Effects Models

The preceding discussion applies to fixed-effects analyses, which assume that sample effect sizes are homogeneous and estimate a *single* population parameter. In contrast, a random-effects analysis assumes that the effect sizes included in a review are heterogeneous and sampled from a *distribution* of population effect sizes (Hedges & Vevea, 1998; Hunter & Schmidt, 2000). Under the random-effects model, we are typically estimating the mean of the distribution of effects and its variance. Thus within the random-effects model, two sources of variability contribute to observed differences among sample effect sizes: (a) sampling error and (b) heterogeneity among the estimated population values. The latter heterogeneity reflects the random-effects variance component ($\hat{\tau}^2$) whose value must be estimated and then added to each individual v_i prior to calculating $\bar{T}.$, the standard error of $\bar{T}.$, and the 95% CI associated with $\bar{T}.$. Adding the random-effects variance component ($\hat{\tau}^2$) to each v_i has two notable consequences: (a) It results in a larger

Table 3
A Cumulative Meta-Analysis of Log Odds Ratios Assessing the Effect of Aspirin on Patient Mortality Following Myocardial Infarction

Study date	$\text{Ln}[or_i]$	n_i	Cumulative n	Cumulative weighted average effect size $\bar{L}.$	95% CI	Power to detect $\text{Ln}[T] = .11$
1974	.329	1,239	1,239	.329	$\bar{L}. \pm .386$	.09
1976	.385	1,529	2,768	.356	$\bar{L}. \pm .277$	.12
1979	.220	1,682	4,450	.289	$\bar{L}. \pm .197$	.20
1979	.222	626	5,076	.280	$\bar{L}. \pm .183$	.22
1980	.226	1,216	6,292	.269	$\bar{L}. \pm .164$	.26
1980	−.125	4,524	10,816	.103	$\bar{L}. \pm .125$	.41
1988	.111	17,187	28,003	.109	$\bar{L}. \pm .065$	.92

Note. Odds ratio data in column 2 are as cited in "The Statistical Basis of Meta-Analysis," by J. L. Fleiss, 1993, *Statistical Methods in Medical Research, 2,* Table 6, p. 134, and Table 7, p. 136. Copyright 1993 by the Division of Biostatistics, Columbia School of Public Health. Adapted with permission. Frequency data are presented in the Appendix. $\text{Ln}[or_i]$ = natural log of the odds ratio; n = sample size; $\bar{L}.$ = the weighted average effect size in the log odds metric; CI = confidence interval; $\text{Ln}[T]$ = natural logarithm of the population odds ratio.

estimate of the standard error of $\bar{T}.$ (i.e., $\sqrt{V}.$), and (b) it results in a wider 95% CI around $\bar{T}.$ (i.e., $\pm$ 1.96 $\times$ ($\sqrt{V}^*.$), where $V^*.$ is the standard error of the mean $\bar{T}.$ under the random-effects model).

Several writers recommend using random-effects (RE) analyses rather than fixed-effects (FE) analyses because RE analyses yield wider confidence intervals around the weighted average effect size, thereby reducing the likelihood of committing a Type I error. From this standpoint RE analyses are regarded as more "conservative" than FE analyses (Lau et al., 1992; see Poole and Greenland, 1999, for an alternative view). Indeed, if effect sizes are heterogeneous then FE analyses can produce Type I error rates that are 10 times greater than the reported .05 alpha level (see, e.g., Hunter & Schmidt, 2000, Table 1, p. 279). In addition to reducing the likelihood of Type I errors, the wider confidence intervals that accompany RE analyses reduce the tendency to attribute greater precision to the estimated effect size than is justified (a problem that may accompany FE analyses). Perhaps most importantly, RE analyses may permit generalizations that extend beyond the studies included in a review, whereas FE analyses are more restrictive and only permit inferences about estimated parameters. The preceding concerns led Hunter and Schmidt (2000) to conclude "that FE models and procedures are rarely, if ever, appropriate for real data in meta-analysis and that therefore RE models and procedures should be used in preference to FE models and procedures" (p. 284).

Despite the increasing use of RE analyses, several limitations associated with RE analyses should be noted. First, the studies included in a meta-analysis probably never represent a random selection of studies that could have been conducted in a research domain, thereby violating a basic assumption of the RE model. Indeed, the studies included in a meta-analysis rarely represent even a random sample of studies that have been conducted in a research domain, again violating an assumption of the RE model. In both cases, RE analyses may yield inferences that are misleading. Perhaps more importantly, the two types of analyses (FE and RE) address fundamentally different research questions, a point that is often misunderstood. Replacing one set of analyses (e.g., FE) with another set of analyses (RE) can introduce an unintentional "sleight of hand" whereby one research question is unknowingly substituted for another. Within an FE model, the research question might be phrased as follows: "What is the best estimate of the population effect size and is it of practical or theoretical importance?" The statistical null hypothesis typically assumes that θ is zero; computing the mean effect size and its associated 95% CI provides a test of this hypothesis.

Within the RE model, however, the goal is not the estimation of a single population effect size because the RE model presupposes a distribution of population effect sizes from which sample effect sizes have been drawn. Thus computing the mean of the distribution of population effect sizes (RE model) yields different information than computing the mean of the distribution of sample effect sizes (FE model). Shadish (1992) and Hedges and Pigott (2001) noted that within the RE model there may be instances in which the average population effect size is positive yet some of the individual population effect sizes may be zero or negative, which "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" (Hedges & Pigott, 2001, p. 211). Likewise, there will be instances in which the mean population effect size is zero although some of the individual population effect sizes are themselves positive (and others negative) in value. The research question underlying an RE analysis might be phrased as follows: "What is the range and distribution of population effect sizes, and what proportion of these values is small or large, negative or positive?"

Statistical Power and the Random-Effects Model

Random-effects meta-analyses also can increase statistical power by reducing the standard error of the weighted average effect size. However, increasing the number of studies in a random-effects meta-analysis from k to $k + 1$ does not always yield increased statistical power (as it does in a fixed-effects analysis). Recall that within the random-effects model two sources of variability contribute to observed differences among sample effect sizes—sampling error and heterogeneity among the estimated population values. Between-studies heterogeneity is reflected in the random-effects variance component ($\hat{\tau}^2$), an estimate of which is added to each individual v_i prior to calculating the standard error of $\bar{d}.$ and its 95% CI. One method for estimating the random-effects variance component ($\hat{\tau}^2$) is provided by Hedges and Vevea (1998): $\hat{\tau}^2 = [Q - (k - 1)]/c$ when $Q \geq k - 1$, where $Q = \sum w_i (T_i - \bar{T}.)^2$ and $c = \sum w_i - [\sum(w_i)^2/\sum w_i]$. If the variance-component estimate derived from a set of

$k + 1$ studies is larger than the variance component estimate derived from the initial set of k studies, then the standard error of $\overline{T}.$ may increase, resulting in decreased statistical power.

One example illustrates this point. The data presented in Table 1 were subjected to a random-effects analysis; the findings are presented in Table 4. The cumulative meta-analysis of the first four studies yields a 95% CI that is larger than the 95% CI derived from the cumulative meta-analysis of the first three studies ($\overline{d}. \pm 0.402$ vs. $\overline{d}. \pm 0.297$, respectively). Indeed, statistical power has decreased from .66 to .42 despite the increased cumulative sample size. The power of the cumulative RE meta-analysis does not rise above .66 (the power based on the first three studies) until seven studies have been included in the review. The decreased statistical power results from the larger variance-component estimates that are added to each v_i associated with the cumulative meta-analyses of four, five, and six studies in the present example.

The above discussion does not imply that fixed-effects analyses are preferable to random-effects analyses. In general, random-effects analyses also increase statistical power compared with that of a single study, although the addition of any particular study to the analysis does not guarantee an increase in power. The decision to use a fixed- or random-effects analysis should depend on other considerations, including the types of inferences that one seeks to make and the heterogeneity of the studies included in a review. Perhaps most importantly, investigators must decide which model (RE or FE) best characterizes the phenomena under study.

Conclusion

The meta-analysis of individual studies increases statistical power by reducing the standard error of the weighted average effect size. A smaller standard error results in a smaller confidence interval around the weighted average effect size, and a small confidence interval increases the likelihood of detecting nonzero population effects. Notably, small confidence intervals also increase the precision of the estimated population effect size, an advantage that is associated with meta-analysis even when the population effect is truly zero (a situation that would preclude discussion of statistical power).

Our strategy for computing the statistical power of studies cited in Tables 1–4 requires clarification. The power of a statistical test is partially a function of the population effect size (θ_1) that one seeks to detect. Because θ_1 is typically unknown, a value for power calculations must be selected on the basis of theory, practical importance, or sample data. In our examples, we adopted the latter strategy, using the weighted average effect size (derived from each meta-analysis in Tables 1–4) as the basis for calculating the statistical power associated with studies in each review. This

Table 4
A Cumulative Meta-Analysis of Sample d_is Assessing Gender Differences in Risk Driving Behaviors (Random-Effects Model)

Study date	d_i	n_1	n_2	Cumulative random-effects variance component	Cumulative weighted average effect size $\overline{d}.$	95% CI	Power to detect $\delta = 0.36$
1978	0.03	28	40	—	—	—	—
1992	0.53	71	118	.0821	0.32	$\overline{d}. \pm 0.483$	.31
1992	0.67	678	673	.0470	0.49	$\overline{d}. \pm 0.297$	.66
1993	0.06	704	672	.1484	0.34	$\overline{d}. \pm 0.402$	.42
1994	0.33	108	125	.1236	0.34	$\overline{d}. \pm 0.331$	.56
1995	–0.08	139	147	.1241	0.27	$\overline{d}. \pm 0.302$	.65
1995	0.75	55	101	.1235	0.33	$\overline{d}. \pm 0.281$	.71
1995	0.48	303	376	.0983	0.35	$\overline{d}. \pm 0.236$	.85
1996	0.61	88	178	.0936	0.38	$\overline{d}. \pm 0.218$	.90
1996	0.22	559	373	.0772	0.36	$\overline{d}. \pm 0.189$	.96
1996	0.23	156	147	.0717	0.35	$\overline{d}. \pm 0.174$	.98
1996	0.43	906	880	.0551	0.36	$\overline{d}. \pm 0.149$	>.99

Note. Sample sizes and effect sizes in columns 3, 4, and 6 are as cited in "Gender Differences in Risk Taking: A Meta-Analysis," by J. P. Byrnes, D. C. Miller, & W. D. Schafer, 1999, *Psychological Bulletin, 125,* Table 1, pp. 374–376. Copyright 1999 by the American Psychological Association. Adapted with permission of author. The random variance component was computed for the first two studies, then the first three studies, and so forth; positive d_i values indicate a male advantage. n_1 = male sample size; n_2 = female sample size; $\overline{d}.$ = weighted average effect size; CI = confidence interval.

strategy should not be confused with a retrospective power analysis in which a researcher uses effect-size data from a single study as the basis for estimating the statistical power associated with the study. Such an approach to power analysis is inappropriate and often leads investigators to mistakenly conclude that their findings would have achieved statistical significance if the power of the test had been greater. In contrast, we used data from multiple studies as a basis for estimating θ_1, which then served as the basis for our retrospective assessment of the statistical power associated with the addition of each study. In a situation where a reviewer wished to compute the power of a proposed meta-analysis, we would urge that a value of θ_1 based on theory or practical impact be used.

The cumulative meta-analyses reported above illustrate the relationships between sample size, meta-analysis summaries, and statistical power; the nature of power in meta-analysis can also be discerned in other cumulative meta-analyses (e.g., Lau et al., 1992). It is important to emphasize, however, that meta-analysis does not inevitably result in identifying nonzero population effects. Indeed, increasing the number of studies in a meta-analysis may yield mean effect sizes that are smaller than initially estimated. Effect-size estimates can be continually revised in response to new information (i.e., studies added to the analysis). Effect-size data presented in Tables 1, 2, and 3 illustrate this point. Mean effect sizes (i.e., $\bar{d}$., $\bar{r}$., and $\bar{L}$.) sometimes decrease in magnitude when a new study is added to the analysis. In contrast, statistical power to detect a specific value of θ_1 always increases following the inclusion of new studies in a fixed-effects meta-analysis, due to smaller standard errors and smaller confidence intervals. The nature of statistical power in random-effects meta-analysis is similar, but not identical, to the situation outlined for the fixed-effects model.

References

Bravata, D. M., & Olkin, I. (2002). Simple pooling versus combining in meta-analysis. *Evaluation and the Health Professions, 24,* 218–230.

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

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, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology, 48,* 71–79.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research, 2,* 121–145.

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

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3,* 339–353.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

Hedges, L. V., & Becker, B. J. (1986). Statistical methods in the meta-analysis of research on gender differences. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through meta-analysis* (pp. 14–50). Baltimore: Johns Hopkins University Press.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88,* 359–369.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: 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., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

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

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8,* 275–292.

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *The New England Journal of Medicine, 327,* 248–254.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

Murlow, C. D. (1995). Rationale for systematic reviews. In I. Chalmers & D. G. Altman (Eds.), *Systematic reviews* (pp. 1–8). London: BMJ Publishing Group.

Poole, C., & Greenland, S. (1999). Random effects models are not always conservative. *American Journal of Epidemiology, 150,* 469–475.

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

Shadish, W. R. (1992). Do family and marital psychotherapies change what people do? A meta-analysis of behavioral outcomes. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, et al. (Eds.), *Meta-analysis for explanation: A casebook.* New York: Russell Sage Foundation.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological), 13,* 238–241.

Strube, M. J. (1985). Power analysis for combining significance levels. *Psychological Bulletin, 98,* 595–599.

Strube, M. J., & Miller, R. H. (1986). Comparison of power rates for combined probability procedures: A simulation study. *Psychological Bulletin, 99,* 407–415.

Zuckerman, M. (1994). *Behavioral expressions and biosocial bases of sensation seeking.* New York: Cambridge University Press.

Appendix

Frequency Data Reported by Fleiss (1993)

Study	Group	Survived	Died
1974	Aspirin	566	49
	Placebo	557	67
1976	Aspirin	714	44
	Placebo	707	64
1979	Aspirin	730	102
	Placebo	724	126
1979	Aspirin	285	32
	Placebo	271	38
1980	Aspirin	725	85
	Placebo	354	52
1980	Aspirin	2,021	246
	Placebo	2,038	219
1988	Aspirin	7,017	1,570
	Placebo	6,880	1,720

Note. Data are from "The Statistical Basis of Meta-Analysis," by J. L. Fleiss, 1993, *Statistical Methods in Medical Research, 2,* pp. 121–145. Copyright 1993 by the Division of Biostatistics, Columbia School of Public Health. Adapted with permission.

Received June 25, 2002
Revision received March 24, 2003
Accepted April 1, 2003 ■