Meta-Analysis

LARRY V. HEDGES and MARTYNA CITKOWICZ

Abstract

Meta-analysis is the use of statistical methods to combine the results of independent research studies. The results of each study are summarized by one or more indices of effect size and a sampling uncertainty (variance) for each effect. Representing study results by effect sizes permits the use of statistical methods to synthesize these results across studies. This essay describes the most frequently used effect sizes and their properties. It describes how the two principal types of analytic methodology in meta-analysis (fixed and random effects models) are used to estimate an average effect across studies. It also discusses how heterogeneity of effects across studies can be detected via a heterogeneity test and modeled as a function of study characteristics. In addition, this essay describes areas of current research in meta-analysis. One area is the development of methods to handle dependencies that can arise when the results of studies are described by several effect sizes computed from data on the same individuals. Another area involves methods for detecting and correcting publication bias. A third is the development of methods to incorporate more complex study designs into metaanalyses, including multilevel experiments and single case designs used in behavioral psychology, special education, and some medicine.

INTRODUCTION

The research literature in many social science fields, such as psychology, economics, education, and political science, has grown rapidly over the last few decades. This has led to the need to organize, summarize, and synthesize findings in a systematic matter. To respond to this need, methods for systematic reviewing of research have emerged (Cooper, Hedges, & Valentine, 2009). One aspect of a systematic review is the analytic step of combining information across studies for the purposes of drawing general conclusions. The use of statistical tools for combining information across studies is called *meta-analysis*.

Meta-analysis represents the results of each study via indices of effect size. Results are summarized across studies using statistical methods to describe a pattern of results. Meta-analysis has emerged as a central tool for integrative analysis in the social sciences, including education and psychology,

Emerging Trends in the Social and Behavioral Sciences. Edited by Robert Scott and Stephen Kosslyn.
ISBN 978-1-118-90077-2.

and in fields as diverse as experimental ecology and medicine. In addition, major systematic efforts have emerged to support the development and dissemination of systematic reviews, including the Cochrane Collaboration in medicine, the Campbell Collaboration in the social sciences, and the What Works Clearinghouse in education.

This essay describes the basic tools for conducting a meta-analysis and outlines some of the major difficulties in completing the meta-analysis (e.g., finding the studies, dealing with publication bias, and modeling dependencies). The goal of this essay is to provide an introduction to the statistical methods used to conduct meta-analyses, and inform readers about the latest developments and issues in the field. It is not meant to be a comprehensive guide to meta-analysis, but rather a useful source for learning about the basic concepts. The readers are encouraged to read some of the references in order to gain in-depth knowledge of the issues presented if they plan to conduct a meta-analysis.

FOUNDATIONAL RESEARCH

Effect Sizes

Effect sizes are numerical indices of study results that represent the findings of a study in a manner intended to be comparable across studies. There are many different effect sizes, but we focus here on effect sizes for studies that compare a treatment group with a control group.

The effect sizes usually used in meta-analysis have standard errors of estimation, which are largely a function of the sample size in the study, and can be computed from analytic formulas. In this section, we describe several effect size indices and show how to compute their sampling variances (the square of their standard errors). The (sample) effect size (estimates) and their variances are the basic inputs required from each study in the meta-analysis.

Studies Measuring Outcomes on a Continuous Scale. If each study evaluates the effect of a treatment by comparing the mean of a treatment group with the mean of a control group and the outcome measurements are normally distributed within the treatment groups with equal variances, the natural effect size parameter is the standardized mean difference (sometimes called *Cohen's d*):

$$\delta = \frac{\mu^{\mathrm{T}} - \mu^{\mathrm{C}}}{\sigma}$$

where the parameters μ^{T} and μ^{C} are the treatment and control group means, respectively, and the parameter σ is the within-group standard deviation. The quantity δ represents the treatment effect in standard deviation units. However, because δ is a population parameter, it is not observed. We use the study sample to estimate or draw inferences about δ. The natural sample estimate of δ is

$$d = \frac{\overline{Y}^{\mathrm{T}} - \overline{Y}^{\mathrm{C}}}{S}$$

where $\overline{Y}^{\mathrm{T}}$ and $\overline{Y}^{\mathrm{C}}$ are the treatment and control group sample means and S is the pooled within-groups standard deviation. This estimate is often modified slightly to produce an unbiased estimate of δ (sometimes called *Hedges's g*):

$$g = d\left(1 - \frac{3}{4\left(n^{\mathrm{T}} + n^{\mathrm{C}}\right) - 9}\right)$$

where n^{T} and n^{C} are the sample sizes in the treatment and control groups of the study and d is the sample standardized mean difference effect.

The variance of g is determined (mostly) by the sample sizes and (slightly) by the magnitude of g. Specifically, the variance, v, of g can be computed as

$$v = \frac{n^{\mathrm{T}} + n^{\mathrm{C}}}{n^{\mathrm{T}} n^{\mathrm{C}}} + \frac{g^2}{2(n^{\mathrm{T}} + n^{\mathrm{C}})}$$

The effect size g is approximately normally distributed with a mean of δ and a variance of v.

Effect Sizes for Other Situations. If both the outcome and independent variables are continuous measures, as in correlational studies, the natural effect size parameter is often ρ, the Pearson correlation coefficient. Its sample estimate is r, the sample correlation.

In order to apply normal theory, we must use a transformation of r, the Fisher z transformation where

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

which is approximately normally distributed with variance $v = 1/(n - 3)$. Here, n is the total sample size in the correlational study. Statistical analyses of correlations (e.g., computing confidence intervals or combining them across studies) are usually carried out in the metric of the z-transform (see, e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009).

When studies measuring outcomes are on a binary scale (such as survival), effect sizes are usually defined in terms of comparisons of the proportion of individuals in the treatment (π^{T}) and control (π^{C}) groups with a particular outcome. The difference ($\pi^{\mathrm{T}} - \pi^{\mathrm{C}}$), ratio ($\pi^{\mathrm{T}}/\pi^{\mathrm{C}}$), or odds ratio $\omega = \pi^{\mathrm{T}}(1 - \pi^{\mathrm{C}})/\pi^{\mathrm{C}}(1 - \pi^{\mathrm{T}})$ of proportions may be used. The odds ratio is generally preferable for the statistical analysis because of its superior statistical properties. For more information about effect sizes based on discrete data, see Fleiss and Berlin (2009); and for a detailed discussion of effect sizes based on continuous data, see Borenstein (2009).

Combining Effect Sizes

Methods for combining estimates of effect size across studies are generally the same, regardless of the effect size index used. Therefore, the methods for meta-analysis using a general effect size parameter are presented, which is denoted by θ, and a general effect size estimate is denoted by T with its variance denoted by v. Thus, the raw data for a meta-analysis of k studies are the effect size estimates $T_1, \ldots, T_k$ and their variances $v_1, \ldots, v_k$. The estimate from the ith study T_i estimates the unknown population effect size parameter θ_i.

The summary of a collection of effect sizes via meta-analysis addresses two basic questions. The first concerns the typical or average value of the effect sizes. The second concerns the consistency of effect sizes across studies. The typical effect size in meta-analyses is estimated by averaging estimates across studies. However, because some studies produce more precise estimates (i.e., they have smaller variances) than others, it makes sense to give more weight to some (the more precise) estimates than others. Two major statistical approaches to meta-analysis differ in how they compute these weights (w's). Fixed effects methods do not include between-study heterogeneity in computing weights, while random effects methods include between-study variation in computing weights, which are described later.

Fixed Effects Methods. If the effect size parameters are identical across studies so that $\theta_1 = \cdots = \theta_k = \theta$, then the most precise estimate of θ is given by the weighted mean effect size:

$$\overline{T}_{\bullet} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}$$

where $w_i = 1/v_i$, so that the weight given to a particular effect size is the inverse of its variance. Because each of the effect size estimates is normally distributed, the weighted mean $\overline{T}_\bullet$ is also normally distributed and the variance $v_\bullet$ of $\overline{T}_\bullet$ is the reciprocal of the sum of the weights

$$v_\bullet = \left(\sum_{i=1}^{k} w_i\right)^{-1}$$

Note that if the θ_i differ, then $\overline{T}_\bullet$ estimates a weighted average of the θ_i's. A 95% confidence interval for θ is given by

$$\overline{T}_\bullet - 1.96\sqrt{v_\bullet} \leq \theta \leq \overline{T}_\bullet + 1.96\sqrt{v_\bullet}$$

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

$$Z = \frac{\overline{T}_\bullet}{\sqrt{v_\bullet}}$$

The level α two-tailed test rejects the null hypothesis when $|Z|$ exceeds the 100α percent critical value of the standard normal distribution (e.g., 1.96 for $\alpha = 0.05$).

The weighted mean provides a summary of the common effect size estimates if they are reasonably homogeneous, but it is important to understand whether the hypothesis that $\theta_1 = \cdots = \theta_k$ is reasonably consistent with the evidence. To test the hypothesis that the effect sizes are the same across studies, we usually use the statistic

$$Q = \sum_{i=1}^{k} w_i(T_i - \overline{T}_\bullet)^2$$

When the effect size parameters are identical, Q has a chi-square distribution with $(k - 1)$ degrees of freedom. Therefore, a test of the null hypothesis that effect sizes are identical across studies at significance level α consists of comparing the obtained value of Q with the upper α critical value of the chi-square distribution with $(k - 1)$ degrees of freedom, and rejecting the null hypothesis of identical effect sizes if Q exceeds this critical value.

Note, however, that this test may not be very powerful when the number of studies included in the analysis is small or if the variances of the effect sizes are large [e.g., if the sample sizes in most studies are small; see Hedges and Pigott (2001)]. Care should be taken when interpreting the results of the test, unless the number of studies is large or they have large sample sizes (so the v_i are small).

Random Effects Methods An alternative method for combining estimates across studies is the random effects model. In this method, studies are considered a sample of possible studies and their effect size parameters are considered a sample from a universe of possible effect size estimates. The objective is to estimate the mean μ and between-study variance τ^2 of the population of effect sizes (the population of θ values) from which the observed study effect sizes are a sample. Note that this differs from the objective in the fixed effects model, which is to estimate effect size (or weighted mean of effect sizes) in the studies that are observed. Thus, the choice of inference model should be governed by the objective of the meta-analysis, rather than by the observed heterogeneity of effects.

If the effect size parameters corresponding to the studies in our sample of studies ($\theta_1, \ldots, \theta_k$) were observed, we could simply compute their variance as a sample estimate of τ^2. Because they are not observed, we must estimate their variance indirectly. We do so by noting that the variance of the observed effect size estimates ($T_1, \ldots, T_k$) depends partly on v_i, which represents estimation errors, and partly on τ^2, which represents true heterogeneity among the θ_i. The Q-statistic used to test heterogeneity is a weighted sample variance that can be used to obtain an indirect estimate of τ^2. In particular,

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

(if the quantity on the right-hand side of the equation is positive, and zero otherwise), where c is a normalizing constant given by

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}$$

Random effects methods compute the weighted mean effect size

$$\overline{T}_{\bullet}^{*} = \frac{\sum_{i=1}^{k} w_i^* T_i}{\sum_{i=1}^{k} w_i^*}$$

where $w_i^* = 1/v_i^* = 1/(v_i + \hat{\tau}^2)$. This corresponds to weighting each effect size by the inverse of the new variance, $v_i^* = v_i + \hat{\tau}^2$, which includes a component of between-study variation τ^2. As in the fixed effect case, the weighted mean

$\overline{T}^*_\bullet$ is also normally distributed, the variance $v^*_\bullet$ of $\overline{T}^*_\bullet$ is the reciprocal of the sum of the weights

$$v^*_\bullet = \left(\sum_{i=1}^{k} w^*_i\right)^{-1}$$

and a 95% confidence interval for μ is given by

$$\overline{T}^*_\bullet - 1.96\sqrt{v^*_\bullet} \le \mu \le \overline{T}^*_\bullet + 1.96\sqrt{v^*_\bullet}$$

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

$$Z *= \frac{\overline{T}^*_\bullet}{\sqrt{v^*_\bullet}}$$

The level α two-tailed test rejects the null hypothesis when $|Z|$ exceeds the 100α percent critical value of the standard normal distribution (e.g., 1.96 for $\alpha = 0.05$).

The fixed- and random-effects weighted means are similar in form and differ only in the weights used to compute them. When $\hat{\tau}^2 > 0$, the $w_i{}^*$ are more similar to one another than the w_i. This means that studies receive more equal weights in the random-effects weighted mean than in the fixed-effects weighted mean. In the latter case, one study can dominate (receive very large weight) if it has a very small variance (usually because it has a very large sample size). In contrast, in the random-effects weighted mean, where the weights given to each study are more similar, no single study can completely dominate. Similarly, when $\hat{\tau}^2 > 0$, each $w_i{}^*$ is larger than the corresponding w_i. Because the variance of the weighted mean is the reciprocal of the sum of the weights, the variance $v^*_\bullet$ of the random-effects weighted mean $\overline{T}^*_\bullet$ is larger than the variance $v_\bullet$ of the fixed-effects weighted mean $\overline{T}_\bullet$. Consequently, confidence intervals for the random-effects weighted mean are longer than those of the fixed-effects weighted mean.

Note that a test of the hypothesis that $\tau^2 = 0$ in the random effects analysis is exactly the test of the hypothesis that $\theta_1 = \cdots = \theta_k$ based on the Q statistics described in connection with the fixed effects analysis, since if $\tau^2 = 0$, then the effect size parameters will be identical.

A quantitative description of the amount of heterogeneity can be provided in either one of two ways. The estimate $\hat{\tau}^2$ of τ^2 provides one such estimate. The square root of this estimate, $\hat{\tau}$, is an estimate of the standard deviation of the distribution of the effect size parameters across studies. An alternative way to characterize heterogeneity is to describe the percentage of variation in the observed effect size estimates that is due to variation in the θ's. The estimate

$$I^2 = \left(\frac{Q - (k-1)}{Q}\right) \times 100\%$$

does exactly that. Because $\hat{\tau}$ describes the *absolute* amount of variation in the θ's and I^2 describes the percentage of variation *relative* to the total variation of estimates (including the amount of variation due to both variation of the θ's and errors of estimation), both are complementary ways to describe variation in effect size parameters.

Modeling Covariates

There are also more elaborate meta-analytic methods for modeling variation across studies as a function of study-level covariates. One style of analyses is designed to determine whether the average effect sizes of subgroups of studies differ from one another, a meta-analytic generalization of the analysis of variance. Another style of analysis examines the relation between continuously measured covariates and effect size, a meta-analytic generalization of the regression analysis (sometimes called *meta-regression*). For more information about the comparison of groups of effects, see Konstantopoulos and Hedges (2009); and for information about meta-regression, see Raudenbush (2009).

CUTTING-EDGE RESEARCH

In the last two decades, the literature on methods for meta-analysis has expanded substantially [see Sutton and Higgins (2008) for a review of recent developments in meta-analysis]. Space does not permit us to review all of this literature, so we focus on those we believe to be most important in this essay. Specifically, we explain ways to model dependencies, define methods that account for publication bias, and outline approaches for dealing with more complex research designs.

Modeling Dependencies

Often studies measure the outcome of a study in more than one way, giving rise to more than one effect size estimate per study, which are not statistically independent of one another. One of the vexing practical problems in meta-analysis arises when the meta-analyst wants to combine information from all of these effect sizes. One approach to this problem is to formally model the dependencies among effect sizes from the same study by specifying the correlations among them and then use multivariate methods (see, e.g., Hedges & Olkin, 1985 or Kalaian & Raudenbush, 1996). Although this approach is elegant, it is difficult to use because the information needed to compute correlations among effect size measures within studies is seldom reported. Even when such information is reported, it is tedious to use.

A new approach to the problem involves the use of empirical variance estimates that do not require information about the correlations among effect size estimates (Hedges, Tipton, & Johnson, 2010). These methods are considerably easier to use and provide valid statistical analyses (significance tests and confidence intervals) even when there may be several correlated effect size estimates from each study. One limitation of these methods is that a moderate to large number of studies are needed (usually 20 or more studies). New research is improving these methods so that they may be used with a much smaller number of studies.

Publication Bias

Statistically nonsignificant findings are less likely to be published than findings that find an effect because they are sometimes viewed as uninteresting or of lesser quality. That leads to a published literature that is unrepresentative of all completed studies and can result in substantial biases, called *publication bias*. Meta-analyses are often biased because unpublished studies are substantially more difficult to find (and include in the analysis) than published studies. Thus, syntheses that underrepresent unpublished studies may tend to report average effects that are larger (in absolute value) than what they would be if all completed studies were included in the analyses because the studies that are missing most likely contain nonsignificant effects (Dickersin, 2005). A number of techniques have been developed to estimate and reduce the impact of publication bias in meta-analysis. We outline the most commonly used and some more sophisticated methods later.

One class of methods is based on the principle that, if there is publication selection based on effect size or statistical significance, there should be a relation between effect size estimates and sample size (or variance). Funnel plots, which are scatterplots of the treatment effect estimate plotted against study sample size (or variance, which is a function of sample size), are often used to make a visual assessment of this relation. An asymmetrical plot is what suggests bias may be present in the meta-analysis. One difficulty in using funnel plots is that, when the number of studies is small, it is not easy to make a visual determination of whether the plot is symmetric or not. Consequently, analytic approaches have been developed in order to quantify funnel plot asymmetry. For example, Begg and Mazumdar's (1994) rank correlation method examines the strength of the association between effect-size estimates and their sampling variances, while Egger's linear regression approach determines *whether* there is a linear relationship between the two estimates (Egger, Smith, Schneider, & Minder, 1997). However, even these methods are not very sensitive when the number of studies is small.

Duval and Tweedie (2000) took a different approach by developing a type of sensitivity analysis that assesses the possible impact of publication bias on meta-analyses. This method, called the *trim-and-fill*, uses funnel plot asymmetry to estimate how many effects might be missing due to publication selection, imputes values for the potentially missing effects, and recalculates the weighted mean effect once the imputed effects are added to the original data set. The method gives an adjusted estimate of the average effect size and thus can provide a quantitative estimate of the potential impact of publication bias on a particular meta-analysis. However, this method assumes that publication bias follows a deterministic pattern: it is always the most extreme points in the tail of the distribution that are assumed to be missing.

Although funnel plot asymmetry (a relation between effect size and sample size) may indicate publication bias, it may also be caused by heterogeneity in effects when there is no publication bias. Therefore, all methods of detecting publication bias based on funnel plot asymmetry have a common weakness. Although this weakness can be addressed if covariates can be found to explain all of the variation between studies, this is often not possible, and even when it is, the use of covariates exacerbates the problem of insensitivity of the methods when the number of studies is small (Peters, Sutton, Jones, Abrams, & Rushton, 2006).

A different approach is to adjust the meta-analytic results for bias using a model of the publication selection process. This approach uses a model with two parts: (i) an effect size model (i.e., the standard meta-analytic model that would be used if bias were not present), and (ii) a selection model that identifies how the distribution of observed effects is changed by the selection process. Several selection models have been proposed, but the most promising one to date is by Vevea and Hedges (1995). Their selection model assumes that the relative probability that an effect is observed depends on its statistical significance. Operationally this probability is specified by a step function giving different weights to different intervals of *p*-values (e.g., 0.00–0.05, 0.05–0.10). There are two main advantages to the selection model approaches: (i) they can be designed to work with heterogeneous data by including a between-study variance component, and (ii) they can incorporate both discrete and continuous moderators, allowing one to distinguish between systematic study differences and publication bias. The methods require a substantial number of effects to model the selection process with much precision, and they are technically more involved than some of the other approaches. However, new research is improving these methods by simplifying selection models and increasing their sensitivity with a small number of studies (e.g., Citkowicz, 2012 uses the continuous beta probability density function as the selection model).

Complex Research Designs

Many meta-analyses synthesize studies that use simpler designs, such as randomized controlled trials; however, as more studies use more complicated designs, the need for meta-analytic methods to synthesize them has become apparent. Research on several of these designs has begun only in the last few years. We outline three of them below.

A design that is very common in education is called a *cluster-randomized design* where entire sites (e.g., schools) are assigned to a treatment or control group. Such designs may not be analyzed using standard statistical methods, as they involve two-stage cluster samples that include a between-cluster variance component (in addition to the within-cluster variance that one would normally calculate) that needs to be accounted for. Hedges addressed this issue in meta-analysis by deriving methods for the calculation of effect sizes when the summary data come from a clustered two-level design (e.g., when students are clustered within classrooms; Hedges, 2007) and from three-level designs (e.g., when students are clustered within classrooms, which are then clustered within schools; Hedges, 2011).

Single case designs are widely used in behavioral psychology, special education, and some medical specialties. They permit the evaluation of treatment effects on one individual over time via repeated measures, using the individual as their own control. Although numerous effect sizes had been proposed for single case designs, there has been no consensus on a "standard" effect size. Recently effect size measures have been developed for single case designs that are rigorously comparable to those used in between-subjects designs and, therefore, permit evidence from single case designs to be included in meta-analyses with effect size data from between-subjects designs. Hedges, Pustejovsky, and Shadish (2012, 2013) derived effect size measures for single case designs comparable to the standardized mean difference, gave formulas for their variances, and demonstrated that they have acceptable statistical properties as effect sizes for meta-analysis.

Repeated measures (or within-subjects) designs are methods that use the same individuals in every condition over an extended period of time. The data are analyzed by estimating growth curve models that examine how the individuals change over time. Vevea and Citkowicz (2010) proposed a method to meta-analyze growth curves from the sample means provided in the studies. As no information about within-subject variance is provided, the method includes a sensitivity analysis in order to attenuate the diagonals of covariance matrix. More work is underway on this and related methods.

KEY ISSUES FOR FUTURE RESEARCH

New methods for meta-analysis are being developed and old methods are modified to deal with new problems that arise. One major advance is the use of Bayesian statistics for meta-analysis. Bayesian methods differ from frequentist, or classical, methods in that both the model parameters and data are considered random, rather than fixed, quantities. This allows one to make probabilistic statements about the distribution of the parameters, which could not be done if working in the classical statistics framework. Moreover, Bayesian methods allow prior knowledge to be incorporated in order to make estimation more efficient or to represent strong subjective beliefs about parameters. The use of Bayesian methods in meta-analysis has gained popularity particularly in medicine [see Sutton and Abrams (2001) for a review of those methods]. Various models have been developed to conduct these meta-analyses using Bayesian statistics; however, more research is needed to expand these models to deal with issues such as publication bias and excess heterogeneity.

An issue that is not discussed often enough is the difficulty of updating meta-analyses with new studies. With 20,000 randomized trials published in PubMed in 2010 alone, it is becoming increasingly harder to keep meta-analyses up to date. In response, Wallace, Trikalinos, Lau, Brodley, and Schmid (2010) developed an online classification tool to semi-automate the screening process. They use machine learning algorithms to screen citations in the biomedical literature. In assessing their method, they found that the number of citations to be screened manually was reduced by 40–50%, with not a single citation eligible for inclusion in a meta-analysis excluded. This new tool will save researchers a lot of time and money; however, it is currently only available in the biomedical field. It would be useful to expand it to the social sciences.

REFERENCES

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101.

Borenstein, M. (2009). Effect sizes in continuous data. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221–235). New York, NY: Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, England: John Wiley & Sons, Ltd.

Citkowicz, M. (2012). *A parsimonious weight function for modeling publication bias*. (Unpublished doctoral dissertation). University of California, Merced, CA.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessments, and adjustments* (pp. 11–33). West Sussex, England: John Wiley & Sons, Ltd.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634.

Fleiss, J. L., & Berlin, J. A. (2009). Effect sizes for dichotomous data. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 237–253). New York, NY: Russell Sage Foundation.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. doi:10.3102/1076998606298043

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*(3), 346–380. doi:10.3102/1076998610376617

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

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Journal of Research Synthesis Methods*, *3*, 224–239.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs. *Journal of Research Synthesis Methods*, *4*, 324–341.

Hedges, L. V., Tipton, E., & Johnson, M. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65 (Erratum, *1*, 164–165).

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*(3), 227–235.

Konstantopoulos, S., & Hedges, L. V. (2009). Fixed effects models. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 279–293). New York, NY: Russell Sage Foundation.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *Journal of the American Medical Association*, *295*(6), 676–680.

Raudenbush, S. W. (2009). Random effects models. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295–315). New York, NY: Russell Sage Foundation.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*, 277–303.

Sutton, A. J., & Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Statistics in Medicine*, *27*, 625–650. doi:10.1002/sim.2934

Vevea, J. L., & Citkowicz, M. (2010, July). *Meta-analysis of growth curves from sample means*. Paper presented at the annual meeting of the Society for Research Synthesis Methodology, Cartagena, Spain.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435.

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, *11*, 55+.

FURTHER READING

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504.

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

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessments, and adjustments*. West Sussex, England: John Wiley & Sons, Ltd.

LARRY V. HEDGES SHORT BIOGRAPHY

Larry V. Hedges is the Board of Trustees Professor of statistics and professor of educational and social policy at Northwestern University. He was formerly the Stella M. Rowley Distinguished Service Professor of education, psychology, and sociology at the University of Chicago. Hedges's research interests include the development of statistical methods for educational and social research, the use of statistical concepts in social and cognitive theory, and social policy analysis. Major areas of his methodological work include the development of statistical methods for meta-analysis and the design and interpretation of social experiments. He is a member of the National Education Sciences Board, the National Academy of Education, a Fellow of the American Academy of Arts and Sciences, the American Educational Research Association, the American Statistical Association, and the American Psychological Association. He was Methods Editor of the *Journal of Research on Educational Effectiveness*, Editor of the *Journal of Educational and Behavioral Statistics*, Quantitative Methods Editor of *Psychological Bulletin*, and Associate Editor of the *American Journal of Sociology*. His books include *Statistical Methods for Meta-analysis* (with Ingram Olkin) and *The Handbook of Research Synthesis and Meta-analysis* (with Harris Cooper and Jeff Valentine).

Academic webpage:
http://www.ipr.northwestern.edu/people/hedges.html.

MARTYNA CITKOWICZ SHORT BIOGRAPHY

Martyna Citkowicz is a Quantitative Researcher at the American Institutes for Research. She received her doctoral degree in psychological sciences with an emphasis in quantitative psychology at the University of California, Merced, in 2012. Citkowicz's research focuses on statistical analyses and solutions for methodological problems in the social sciences. Most of her work has centered on assessing and developing methods in meta-analysis, including examining random- and fixed-effects inference, publication bias, variance component estimation, and growth curve modeling. She has written numerous essays on the topic and presented her research at both national and international conferences.

Academic webpage:
http://northwestern.academia.edu/MartynaCitkowicz.

RELATED ESSAYS

Social Epigenetics: Incorporating Epigenetic Effects as Social Cause and Consequence *(Sociology)*, Douglas L. Anderton and Kathleen F. Arcaro

To Flop Is Human: Inventing Better Scientific Approaches to Anticipating Failure *(Methods)*, Robert Boruch and Alan Ruby

Repeated Cross-Sections in Survey Data *(Methods)*, Henry E. Brady and Richard Johnston

Ambulatory Assessment: Methods for Studying Everyday Life *(Methods)*, Tamlin S. Conner and Matthias R. Mehl

Models of Nonlinear Growth *(Methods)*, Patrick Coulombe and James P. Selig

Quantile Regression Methods *(Methods)*, Bernd Fitzenberger and Ralf Andreas Wilke

The Evidence-Based Practice Movement *(Sociology)*, Edward W. Gondolf

The Use of Geophysical Survey in Archaeology *(Methods)*, Timothy J. Horsley

Network Research Experiments *(Methods)*, Allen L. Linton and Betsy Sinclair

Longitudinal Data Analysis *(Methods)*, Todd D. Little *et al.*

Structural Equation Modeling and Latent Variable Approaches *(Methods)*, Alex Liu

Data Mining *(Methods)*, Gregg R. Murray and Anthony Scime

Remote Sensing with Satellite Technology *(Archaeology)*, Sarah Parcak

Quasi-Experiments *(Methods)*, Charles S. Reichard

Digital Methods for Web Research *(Methods)*, Richard Rogers

Virtual Worlds as Laboratories *(Methods)*, Travis L. Ross *et al.*

Modeling Life Course Structure: The Triple Helix *(Sociology)*, Tom Schuller

Content Analysis *(Methods)*, Steven E. Stemler
Person-Centered Analysis *(Methods)*, Alexander von Eye and Wolfgang Wiedermann
Translational Sociology *(Sociology)*, Elaine Wethington