CHAPTER 10

Meta-analysis

DAVID B. WILSON

Advancement in science depends on a clear understanding of what is currently known. A challenge in many areas of science, including criminology and criminal justice, is making sense of rather disparate findings across studies of a common research question and grappling with a large number of studies. Meta-analysis is a statistical method designed to tackle these problems and approaches the task of taking stock of the literature as a research endeavor in its own right. That is, meta-analysis applies the methods and logic of social science research to the task of synthesizing results across studies and exploring explanations for variability in those results.

The logic of meta-analysis is straightforward and examples date back over 100 years. Arguably, the earliest was by Karl Pearson, the developer of the Pearson's product moment correlation coefficient (Pearson 1904, as cited in Hunt (1997)). Pearson synthesized the finding from multiple studies of the effectiveness of inoculation for typhoid fever. The very idea of inoculations was controversial among the medical community at the time. When looked at individually, the results for the effectiveness of inoculation were inconsistent, with some studies finding a statistically significant effect and others not. Pearson computed the tetrachoric correlation between inoculation and mortality within each study and then averaged the correlations across studies. The average result across studies clearly supported the value of inoculations. By today's standard, this was a meta-analysis, although the term was not introduced until the 1970s (Glass 1976) and the statistical methods have undergone substantial refinement.

Meta-analysis addresses a primary complication of synthesizing results across studies: findings will differ. As an example, imagine that you are interested in the effectiveness of a new treatment program, called XYZ, for adjudicated juveniles. An evaluation randomly assigned 200 youth to either the XYZ program or a control condition and found a statistically significant lower level of postprogram delinquent behaviors in the XYZ condition relative to the control condition (43% vs. 57%, respectively). This single study is seen as a promising evidence of the effectiveness of this program. Unfortunately, a second independent evaluation with a smaller sample size, 50 in each condition, failed to find a statistically significant effect. From a traditional perspective, this second study weakens the inference that this program is effective. However, shifting the focus from statistical significance to the magnitude and direction of effect shows that the second study also observed an association between assignment to the XYZ condition and postprogram delinquency (30% vs. 45% for the XYZ

A.R. Piquero and D. Weisburd (eds.), *Handbook of Quantitative Criminology*,
DOI 10.1007/978-0-387-77650-7_10, © Springer Science + Business Media, LLC 2010

and control conditions, respectively). Furthermore, the size of the effect is larger in the second study. When combined with the first study, the results from the second study strengthens the inference that the XYZ program is effective, rather than weakens it.

As seen in this example, the focus in meta-analysis shifts from statistical significance, a common emphasis in more traditional methods of reviewing studies, to the direction and magnitude of observed effects. This avoids a profound limitation of statistical significance: A statistically significant finding is a strong conclusion, whereas a statistically nonsignificant finding is a weak conclusion. Recall that null hypothesis significance testing dictates that we *fail to reject* (not *accept*) the null hypothesis when the p-value is not significant. A nonsignificant finding may simply reflect a lack of sufficient statistical power, a common problem in criminological and criminal justice research (Bushway et al. 2006; Lipsey et al. 1985; Weisburd et al. 2003). As such, statistical significance can be misleading when used as the basis for drawing inferences across a collection of related studies.

For meta-analysis to make sense, the collection of studies on which it is based must be estimating a common relationship of interest, such as that between program XYZ and recidivism or between race and the likelihood of arrest in a police–citizen encounter. If the studies are a collection of pure replications, then the idea of creating an overall estimate of the effect is indisputable, as long as one accepts the logic and assumptions of standard statistical practices in the social sciences. On the basis of these assumptions, we would expect some replications to *overestimate* the true population effect and other replications to *underestimate* the true population effect. The overall average should provide a more accurate and robust estimate of the true population effect.

Collections of pure replications are rare within the social sciences. More typically, studies addressing a common research question will vary with respect to the operationalization of the constructs, the implementation of any experimental manipulation, and other methodological or substantive features. For meta-analysis to be credible, one must be able to argue that the studies are at least *conceptual* replications – each is examining a common empirical relationship despite substantive and methodological variations. The greater the variability in study features, the more abstract the nature of the conceptual replication. For example, the meta-analytic work of Lipsey (e.g., Lipsey 1995; Lipsey and Wilson 1998) on the effectiveness of juvenile delinquency includes great variability in the nature of the intervention. However, at a conceptual level, all of the studies are examining the relationship between a juvenile delinquency intervention and future delinquent behavior.

The analysis of effect-sizes across a collection of pure replications focuses on estimating the common or mean effect-size. This focus shifts to an examination of the relationship between study characteristics and observed effects as the collection of studies moves from pure replications to more abstract conceptual replications. The differences between the studies, both substantive and methodological, and how these differences related to effect-sizes takes on greater meaning than simply the overall mean effect-size.

OVERVIEW OF META-ANALYTIC METHODS

There are several distinct tasks involved in conducting a meta-analysis (see Cooper 1998; Lipsey and Wilson 2001a). The first task relates to problem formulation and involves an explication of the research question(s) or objectives of the meta-analysis. Second, an explicit set of inclusion and exclusion criteria must be specified that clearly define the characteristics of studies that will be included in the meta-analysis. Third, a comprehensive search for all eligible studies, published or unpublished, is conducted. This typically involves searching multiple sources, including bibliographic databases, reference lists of prior reviews and eligible studies, Internet searches, and contacting authors active in the area (see Cooper 1998; Lipsey and Wilson 2001a; Wilson 2009). The fourth task involves the coding of eligible studies. Using a coding form similar to a survey, information about the features of the studies are captured and effect-sizes are computed. The latter represent the results or findings of the studies and are discussed in more detail below. Fifth, the effect-sizes are analyzed using statistical methods specific to meta-analysis. These may include analyses that examine the relationship between coded study features and effects sizes. And finally, the results are interpreted and written up. The focus of this chapter is on the statistical methods of meta-analysis, that is, the computation of effect-sizes and their analysis. Before introducing effect sizes, I discuss issues related to primary study design and meta-analysis.

BASIC RESEARCH DESIGNS

Meta-analysis can be applied to many different research designs. Typically, a single meta-analysis will be focused on studies that share a common research design or at least designs that are conceptually the same and lend themselves to a common effect-size index, such as experimental and quasiexperimental designs with a comparison group. Research designs can be broadly conceptualized as univariate, bivariate, and multivariate. The former are designs estimating a statistical parameter of a single variable, such as a mean or proportion. For example, with colleagues I have conducted a meta-analysis of the proportion of homicide victims testing positive for illicit drugs (Kuhns et al. 2008).

Bivariate research designs fall into two main types, those examining a correlation between naturally occurring variables and those examining the relationship between an experimental (or potentially manipulated) variable and a dependent variable. An example of the former type is a meta-analysis by Lipsey and Derzon (1998) of cross-sectional and longitudinal studies examining the correlation between risk factors and delinquent behavior. Examples of the latter are bivariate designs examining the effectiveness of correctional programs, police activities, or other policies designed to reduce crime. Lipsey and Cullen (2007) provide a recent review of such meta-analyses within criminology and criminal justice.

Multivariate research involves the examination of three or more variables. Although there are examples of meta-analyses of such research, often the focus is on a specific bivariate relationship imbedded within the multivariate context. For example, Pratt and Cullen (2000) examined the relationship between low self-control and crime from a collection of multivariate studies. Later in this chapter, I discuss complications involved in meta-analyzing multivariate research. The analysis of univariate and bivariate research designs is generally straightforward. The first step in the process is the selection of the appropriate effect-size index.

THE EFFECT-SIZE

The key building block of meta-analysis is the effect-size. An effect-size is a statistical measure of the effect of interest that can be computed and compared across studies. In this context, I am using the term *effect-size* in a generic sense – it can refer any statistical index that is aggregated across studies. Common effect-sizes include the standardized mean difference, the correlation coefficient, the odds-ratio, and the risk ratio. The choice of an effect-size type should be driven by the nature of the relationship of interest.

Most meta-analytic work in the social sciences has thus far focused primarily on bivariate relationships although examples of meta-analyses of multivariate relationships and of single variable point-estimates can be found. For a bivariate relationship, the selection of the effect-size type will depend on the nature of both the independent and dependent constructs of interest and whether these are inherently dichotomous (binary) or continuous.

The Standardized Mean Difference

The standardized mean difference (d) is applicable to research designs that involve the comparison of two groups on one or more dependent variable. These two groups may be experimental in nature, such as a treatment group vs. a control group, or naturally occurring, such as boys vs. girls. As the names implies, the effect-size is based on the difference between the means. Thus, this effect-size is best suited to a dependent variable with a continuous underlying construct that will typically be measured in a manner consistent with the computation of means and standard deviations (below I will address how to estimate a standardized mean difference from studies that measure the construct of interest dichotomously). The basic equation for the standardized mean difference (d) is

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}}, \qquad (10.1)$$

where $\overline{X}_1$, s_1^2, and n_1, are the mean, variance, and sample size for group 1, and $\overline{X}_2$, s_2^2, and n_2, are the mean, variance, and sample size for group 2. The denominator of this equation is pooled within groups standard deviation and serves to standardize the difference between the means in terms of the natural variability on this dependent variable, less any variability due to the treatment or group effect. A d of 0.5 indicates that group 1 is half a standard deviation above group 2. It is this standardization that allows for comparison of effects across studies.

Hedges and Olkin (1985) showed that the standardized mean difference effect-size is upwardly biased when based on small sample sizes. While this bias is relatively small when sample sizes exceed 20 (Hedges 1981), it has become standard practice to apply this adjustment to d even if your sample sizes all exceed 20. The equation for adjusting the effect-sizes is

$$d' = \left[1 - \frac{3}{4N - 9}\right] d, \qquad (10.2)$$

where d is from (10.1). This adjusted d is referred to as the *unbiased standardized mean difference* effect-size.

Not all authors report means, standard deviations, and sample sizes for all outcome measures. This necessitates computing d based on other available information. Some of these alternative formulas are algebraically equivalent to (10.1) above, whereas others provide a reasonable estimate. An algebraically equivalent equation for computing d based on a t-value is

$$d = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}}, \qquad (10.3)$$

where t is from an independent t test assessing the difference between two means and n_1 and n_2 are the respective sample sizes. Additional equations for computing d can be found in Borenstein (2009) and Lipsey and Wilson (2001a). A simple effect size calculator is also available at `http:\\mason.gmu.edu\˜dwilsonb\ma.html`.

Another common problem in the application of the d-type effect-size is that some of the studies will use a dichotomous indicator for the dependent variable of interest. For example, a meta-analysis of school-based bullying programs may be interested in aggression as the outcome and aggression may be measured on an interval scale in some studies and dichotomously (aggressive, not aggressive) in others. It is desirable to calculate an effect-size that is comparable across both sets of studies. There are several ways to estimate a d-type effect-size from 2 by 2 data (i.e., treatment vs. control by a dichotomous dependent variable). Computer simulations by Sánchez-Meca et al. (2003) suggest that the best method is the Cox method (Cox 1970), which is based on the logged odds-ratio. Using this method, d is estimated as

$$d = \frac{\ln(\mathrm{OR})}{1.65}, \tag{10.4}$$

where OR is the odds-ratio (see below). This method is very similar to the Hasselblad and Hedges (1995) method which divides the logged odds-ratio by $\pi/\sqrt{3}$ or 1.81. The latter method slightly underestimates d, particularly for large values of the logged odds-ratio, although in practice the difference tends to be slight. The Cox method also produces values that are very close to the probit method.

The Correlation Coefficient

Correlational research is common in the social sciences. The correlation coefficient is a natural effect-size for literature of this type. For example, Lipsey and Derzon (1998) conducted a large meta-analysis of both cross-sectional and longitudinal predictors of delinquency. The dependent variable, delinquency, was conceptualized as continuous in nature and often, but not always, measured on a scale with a range of possible values. Similarly, the independent variables, the risk factors, were also often measured in a continuous fashion or at least in a manner consistent with the computation of a Pearson correlation coefficient.

Extracting correlation coefficients (r) from studies where the correlation is the natural effect-size is generally straightforward: the majority of studies will report the correlation coefficient. However, studies will be identified for which this is not the case. It is often possible to use other available data to compute the correlation coefficient, or at least a close approximation. For example, a study might simply report the t-value associated with a significance test of the correlation coefficient. In this case, r can be computed as

$$r = \frac{t}{\sqrt{t^2 + \mathrm{df}}}, \tag{10.5}$$

where t is the t-value and df is the degrees of freedom. Similarly, if the study only reports the exact p-value for the t-test of the correlation, the t-value can be determined using an inverse distribution function (this is available in most all spreadsheets and statistical software packages). Other equations for computing r can be found in Borenstein (2009) and Lipsey and Wilson (2001a).

The Odds-Ratio and Risk-Ratio

Dichotomous dependent variables are common in criminology and criminal justice. Research designs that examine the relationship between a dichotomous independent variable, such as assignment to a boot-camp vs. prison, and a dichotomous dependent variable, such as arrest within 12-months of release, are well suited to the odds-ratio or risk ratio as the effect-size. Data such as this can be represented in a 2 by 2 contingency table.

The odds ratio is the odds of success (or failure) in one condition relative to the odds of success (or failure) in the other. An odds is the probability of an event relative to its complement, the probability of the absence of an event. For example, assume that 54 of 100 offenders released from a correctional boot-camp were arrested in the first 12-months. Using this data, the probability of an arrest in the boot-camp condition is 0.54 and the probability of not being arrested is 0.46. The odds of an arrest in the boot-camp condition is 0.54/0.46 or 1.17. Assume also that we have a prison condition and that 50 of 100 offenders released from prison were arrest in the first 12-months. The odds of an arrest for the prison condition is $0.50/0.50 = 1$. The ratio of these two odds is the odds ratio, or 1.17/1.00 or 1.17. Thus, the boot-camp condition has a slightly higher odds of an arrest than the prison condition. If we had defined the event of interest as *not* arrest, then the odds ratio would have been the inverse of this or $(0.46/0.54)/(0.50/0.50) = 0.85$. An odds ratio of 1 indicates a null effect. There is a simple way to compute the odds-ratio (OR) using frequencies:

$$\mathrm{OR} = \frac{ad}{bc}, \tag{10.6}$$

where a, b, c, and d are the cell frequencies of a 2 by 2 contingency table. As done above, the odds ratio can also be computed from the proportion exhibiting the event within each condition as

$$\mathrm{OR} = \frac{p_1/\,(1-p_1)}{p_2/\,(1-p_2)}, \tag{10.7}$$

where p_1 is the event rate (proportion) for the first condition, and p_2 is the event rate for the second condition.

The risk ratio (also called relative risk) can also be used in the above situation and is easier to interpret as it is the ratio of the probabilities of the event, rather than the ratio of the odds of the event. For example, the probability of an arrest for the boot-camp condition is 0.54 and the probability of an arrest in the prison condition is 0.50. Thus, the risk ratio is 0.54/0.50 or 1.08. This can be interpreted as an 8% increase in the probability of arrest for the boot-camp condition. Using cell frequencies, the risk ratio (RR) can be computed as follows:

$$\mathrm{RR} = \frac{a/(a+b)}{c/(c+d)} = \frac{p_1}{p_2}, \tag{10.8}$$

where a, b, c, and d are defined as above and p_1 and p_2 are the probability of an event for each group, respectively. Despite its intuitive appeal, the risk ratio has two statistical limitations (Fleiss and Berlin 2009). The first is that when the probability of the event is high in the comparison condition, the theoretically possible values of RR are constrained. This may produce (or add to) heterogeneity in effects across studies. The second is that the two probabilities on which it is based cannot be properly estimated in retrospective studies that select cases based on the outcome (Fleiss and Berlin 2009) (i.e., case-control studies). These designs are common in epidemiological research. A criminological example would be to obtain a sample of delinquent and nondelinquent youths and then compare these groups on past history of potential risk factors. In this design, it is not possible to estimate the probability of delinquency given the presence of the risk factor as cases were selected because they were delinquent. The odds ratio does not suffer from these weaknesses and has other advantages (see for a full discussion Fleiss and Berlin (2009) and Fleiss (1994)).

Other Effect Sizes

The standardized mean difference, correlation coefficient, odds ratio, and risk-ratio are by no means the only possible effect-sizes for meta-analysis. Any statistical parameter that is comparable across studies and reflects the direction and magnitude of the effect of interest can be used as an effect-size. It is important, however, that the statistical parameter used is *not* a direct function of sample size, such as a t-value, and that it has a computable standard error. The latter is critical for determining the inverse-variance weight used in the analysis of effect-sizes.

There are several examples of other effect-sizes that have been used in meta-analysis. One example is the unstandardized mean difference. This would be used in situations in which all of the studies of interest used the same dependent variable, such as the asocial tendencies subscale of the Jesness Inventory. In this situation, there is no need to standardize the mean difference prior to aggregation. An unstandardized mean difference may be more meaningful as it maintains the original metric of the outcome measure. The unstandardized mean difference effect-size index was used in a meta-analysis of the effectiveness of teenage pregnancy prevention programs conducted by Frost and Forrest (1995). The standardized gain score is another example of an alternative effect-size index. There are numerous examples of meta-analyses in the educational literature examining prepost gains across studies. Another example of an alternative effect-size index is a proportion or rate. With colleagues (Kuhns et al. 2008), I conducted a meta-analysis that examined the proportion of murder victims testing positive for illicit substances. For statistical reasons (see Lipsey and Wilson 2001a), the analyses were performed on the logit of the proportion, but the purpose here is simply to illustrate that the effect-size index used in a meta-analysis should fit the research question and type of research design being synthesized.

WORKING EXAMPLE

A sample of studies from a meta-analysis that I conducted with colleagues (Wilson et al. 2005) is used as a working example to illustrate the methods of meta-analysis throughout this chapter. This meta-analysis examined the effectiveness of group-based cognitive-behavioral treatment for adult offenders. Table 10.1 presents the data and odds-ratio effect-size for ten studies included in that meta-analysis. These studies evaluated either the *Moral Reconation* program, or the *Reasoning and Rehabilitation* program. Only those studies with a randomized design or with a strong quasiexperimental design are shown. The latter studies had a credible comparison group and provided evidence of the comparability of the treatment and comparison conditions. Only one effect-size per study is shown in this table. The selected effect-size was based on a decision rule that gave preference to general measures of criminal behavior that were dichotomous in nature. Furthermore, measures of arrest were preferred over measures of

TABLE 10.1. Odds-ratio effect sizes for group-based cognitive-behavioral programs for adult offenders

	Sample Size		% Recidivating					
Author	Treatment	Control	Treatment	Control	Odds-Ratio	Logged OR	*v*	*w*
Burnett (1996)	30	30	0.100	0.200	2.25	0.81	0.579	1.727
Johnson and Hunter (1995)	47	51	0.255	0.294	1.22	0.20	0.206	4.843
Little and Robinson (1989)	115	65	0.200	0.276	1.52	0.42	0.131	7.614
Little et al. (1991)	70	82	0.610	0.700	1.49	0.40	0.118	8.466
Little et al. (1994)	1,052	329	0.655	0.779	1.86	0.62	0.022	45.742
Porporino et al. (1991)	40	23	0.450	0.521	1.33	0.28	0.275	3.633
Porporino and Robinson (1995)	550	207	0.150	0.160	1.08	0.08	0.050	19.919
Robinson (1995)	1,746	379	0.212	0.248	1.25	0.20	0.018	56.895
Ross et al. (1988)	22	23	0.182	0.696	10.29	2.33	0.511	1.958

Note: These studies are a subset of studies included in Wilson et al. (2005) and represent two specific treatment programs (Moral Reconation and Reasoning and Rehabilitation) and studies that were randomized or used high quality quasiexperimental designs

conviction, and measures of conviction were preferred over measures of reinstitutionalization. The first available posttreatment effect-size was used to increase comparability across studies. This decision rule resulted in a single effect-size per study. This differs slightly from the analyses presented in Wilson et al. (2005), in which the analyses were based on a composite effect-size per study, rather than a single selected effect-size.

Examining the data for the first study listed in the table, Burnett (1996), shows that the treatment and control conditions had 30 individuals each. Only 10% of the treatment group recidivated during the follow-up period compared to 20% in the control group. Using equation 10.7, the odds-ratio is computed as

$$\mathrm{OR} = \frac{0.10/\,(1-0.10)}{0.20/\,(1-0.20)} = 0.444.$$

In this meta-analysis, we wanted to have larger values (those greater than 1) associated with a positive treatment effect. This is accomplished easily by taking the inverse of the odds ratio. Inverting the odds ratio simply changes the direction of the effect. For the above odds ratio, the inverse is $1/0.444 = 2.25$. Thus, the treatment group had 2.25 times the odds of success as the control group. The remaining columns in Table 10.1 are discussed below.

META-ANALYSIS OF EFFECT-SIZES

In meta-analysis, the effect-size is the dependent variable and study characteristics are potential independent variables. The starting point of most meta-analyses is an examination of the central tendency of the effect-size distribution: the mean effect-size. Also of interest is the variability in effects across studies. A collection of effect-sizes might be homogeneous or heterogeneous. A homogeneous collection of effect-sizes varies no more than would be expected due to sampling error alone. Essentially, the studies in such a collection are telling a consistent story with respect to the underlying relationship of interest. A heterogeneous collection of effect-sizes reflects genuine differences in the underlying effect being estimated by the studies. More simply, there are real differences in effects across studies. These differences can be explored through moderator analyses that examine whether study characteristics are associated with effect-sizes, or stated differently, account for some of the variability in effects. A moderator analysis might simply compare the means of the effect-sizes across a categorical variable, such as program type, or may adopt a regression based approach with one or more independent variables.

There are two main statistical approaches to estimate the mean effect-size and related statistics. The first is the inverse-variance weight method developed by Hedges and Olkin (1985). This approach is widely used and is broadly applicable to a range of research questions. The second approach is the Hunter and Schmidt method (Hunter and Schmidt 1990, 2004; Schmidt et al. 2009). The Hunter and Schmidt method was developed in the context of validity generalizability research within the area of industrial–organizational psychology. The principle conceptual difference between these approaches is that the Hunter and Schmidt method corrects effects sizes for methodological artifacts, including error of measurement (unreliability) in both the independent and dependent variables, dichotomization of a continuous independent or dependent variable, measurement invalidity, and range restriction. All of these corrections increase the observed effect-size and attempt to estimate the true underlying effect-size given perfect measurement, etc. Unfortunately, in many areas of research, particularly within criminology and criminal justice, the information needed to fully implement the Hunter and Schmidt method is not available (e.g., reliability and validity coefficients), limiting the applicability of the method. However, the Hunter and Schmidt method has been widely used to synthesize psychometric research and is also popular within social psychology. Schulze (2004) provides a nice comparison of these approaches, including Monte Carlo simulations establishing the strengths and weaknesses of each. This chapter will focus on the inverse-variance weight method.

Independence of Effects

A common complication of conducting a meta-analysis is that multiple effect-sizes can be computed from an individual study. These multiple effect-sizes cannot be treated as independent estimates of the effect of interest – they are based on the same sample and as such are statistically dependent. Treating them as independent estimates would result in an overstatement of the precision of the overall meta-analytic results. Thus, it is important to maintain statistical independence among the effect-sizes included in any given analysis.

There are several methods for addressing this issue. First, distinct measurement constructs can be analyzed separately. For example, a meta-analysis of school-based drug-use prevention programs could meta-analyze the effect-sizes based on measures of knowledge, attitudes, and drug-use behavior separately. If multiple effect-sizes remain within a construct category, then there are three options (1) compute the average effect-size (or select the median) from each study, (2) select an effect-size based on a decision rule, or (3) randomly select one effect-size per study. You may also run multiple analyses based on different decision rules. The basic idea is to make sure that only one effect-size per independent sample is included in any given statistical aggregation.

Another alternative is to model the statistical dependencies directly, thus allowing for the inclusion of multiple effect-sizes per study in a given analysis. Methods have been developed for meta-analysis to do this (e.g., Gleser and Olkin 1994; Kalaian and Raudenbush 1996). Unfortunately, these methods are difficult to implement giving currently available software and generally require information not typically reported by authors, such as the correlation among different measures. However, there are situations, such as the examples provided by Gleser and Olkin (1994) and Kalaian and Raudenbush (1996), where the application of these methods is worthwhile.

Weighting of Effect Sizes

Effect-sizes based on larger samples are more precise than effect-sizes based on smaller samples, all other things being equal. For example, a correlation coefficient between impulsivity and aggressive based on a sample size of 200 is a more precise estimate of this relationship than one from a study with a sample size of 50. Intuitively, it makes sense to give greater weight to the more precise estimate. Although sample size would seem like the natural choice of a weight, and was used as a weight in many meta-analyses conducted in the late 1970s and 1980s when the statistical methods of meta-analysis were undergoing rapid development, a more precise statistical indicator of the precision of an effect-size is its standard error. The smaller a standard error, the more precise the effect-size, at least in terms of sampling error. Because greater weight is to be given to the effect-sizes with smaller standard errors, we need a value that is the inverse of the standard error. Hedges and Olkin (1985) showed, however, that the best weight from a statistical standpoint is based on the squared standard-error, or the inverse of the variance.

For some effect-size types, such as the correlation, odds ratio, and risk ratio, the effect-size must be converted into an alternate form for analysis to allow for the computation of the inverse-variance weight. For example, the correlation coefficient does not have an easily computable standard error, but a Fisher's Zr transformed correlation coefficient has an easily computable one. Thus, when analyzing correlation coefficients, the correlation is first converted into a z as follows[1]:

$$z = 0.5 \ln \left(\frac{1+r}{1-r} \right), \tag{10.9}$$

where r is the correlation effect-size. For values of r less than 0.30, the transformation is slight ($r = 0.30$ converts to a $z = 0.31$) but increases as the value of r approaches 1 ($r = 0.90$ converts to a $z = 1.47$). The variance of z is a simple function of sample size:

$$v_z = \frac{1}{n-3}, \tag{10.10}$$

where n is the sample size for the correlation. The inverse of this is $n - 3$. As such, the weight for a z-transformed correlation is essentially the sample size. You can convert final

[1] There is debate within the meta-analytic literature on the relative merits of analyzing the correlation in its raw form or using the Fisher z transformed value (see Field 2001; Hunter and Schmidt 2004). Computer simulations have shown that the raw correlation is slightly downwardly biased but to a lesser degree than the upward bias of the z transformed value. The original purpose, however, of the z transformation was to provide a computable standard error. An alternative approach is to use the raw correlation as the effect-size and approximate the variance as $v = \left(1 - r^2\right)^2 / (n-1)$ (e.g., Hunter and Schmidt 2004; Shadish and Haddock 2009).

meta-analytic results, such as the mean and confidence interval, back into correlations using the following formula:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}. \tag{10.11}$$

As with the correlation, the odds ratio in its raw form does not have a computable standard error. This stems from the asymmetric nature of the odds ratio. Values greater than 1 up to infinity indicate an increased odds whereas values less than 1 but greater than 0 indicate a decreased odds of the event for the target condition relative to the control condition. An odds ratio of 2 reflects the same strength of association as an odds-ratio of 0.5. Similarly, an odds ratio of 4 reflects the same strength of association as an odds ratio of 0.25. The solution to this problem is the take the natural log of the odds ratio as the effect-size (Fleiss and Berlin 2009; Fleiss 1994). This centers the effects around zero (the natural log of 1 is 0) and positive and negative effects are symmetrical (e.g., the natural log of 2 is 0.69 and the natural log of 0.5 is −0.69). The variance of the logged odds ratio is

$$v_{\ln(\mathrm{OR})} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}, \tag{10.12}$$

where a, b, c, and d are defined as above for the odds ratio. The inverse of this value is used as the weight in meta-analyses of logged odds-ratios. Final results can be converted back into odds-ratios through simple exponentiation (e.g., e^x where x is the meta-analytic mean logged odds-ratio or lower and upper bounds of the confidence interval).

The risk-ratio has the same complications as the odds-ratio with the same solution. Meta-analysis is performed on the natural log of the risk-ratio. The variance of the logged risk-ratio is

$$v_{\ln(\mathrm{RR})} = \frac{1 - p_1}{n_1 p_1} + \frac{1 - p_1}{n_1 p_1}, \tag{10.13}$$

where p_1 and p_2 are the proportion of positive events in groups 1 and 2, and n_1 and n_2 are the respective sample sizes (Fleiss and Berlin 2009). As with the logged odds-ratio, final results can be converted back into risk-ratios through exponentiation.

The standardized mean difference effect-size does have a computable standard error and therefore is analyzed in its raw form. The variance of the standardized mean difference is computed as

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2\,(n_1 + n_2)}, \tag{10.14}$$

where d is the small sample size adjusted standardized mean difference, and n_1 and n_2 are the sample sizes for each group.

The formulas for computing the variance, and thus the inverse-variance weight, for other effect-size types can be found in Lipsey and Wilson (2001a). What should be clear at this point is that for meta-analysis to be possible on a collection of research studies of interest, an effect-size index that is comparable across studies and that has a computable variance is required. The statistical methods presented below are generic and suitable to all effect-size types. ES will be used to denote an effect-size, v the variance of the effect-size, and w the inverse-variance weight (i.e., $1/v$).

Table 10.1 presents the logged-odds ratio, variance, and inverse-variance weight for our example meta-analysis. Focusing on the first row of data, the natural log of the odds-ratio is 0.81. Note that had we left the odds-ratios in their natural direction, with values less than 1 indicating a positive treatment effect (i.e., a reduction in recidivism), then the logged odds-ratio for the calculated odds-ratio of 0.444 is −0.81. This clearly shows that taking the inverse of an odds-ratio simply changes the direction of the effect. What is critical in a meta-analysis is that you code effect-sizes so that the direction of the effect-sizes maintains a consistent meaning across studies. The variance of the logged odds-ratio is computed using (10.12). The cell frequencies of the 2 by 2 contingency table are not provided in Table 10.1. They can, however, be determined through simple algebra. For the treatment group, the number of individuals recidivating (cell a) is the sample size times the percent recidivating. The remaining cell frequencies can be found in a similar fashion. Applying (10.12) to the first row in Table 10.1 produces a variance of

$$v = \frac{1}{3} + \frac{1}{27} + \frac{1}{6} + \frac{1}{24} = 0.579.$$

The inverse-variance weight is simply the inverse of this value:

$$w = \frac{1}{0.579} = 1.727.$$

Fixed-Effects and Random-Effects Models

In conducting a meta-analysis, you must choose between a fixed- or random-effects model. The two models make very different assumptions about the data. The fixed-effects model assumes that all of the studies are estimating a common population effect-size and that the differences between the studies are simply a function of sampling error (Overton 1998; Lipsey and Wilson 2001a; Hedges and Olkin 1985). This is rarely plausible with the possible exception of a meta-analysis of pure replications. In the case of moderator analyses, the fixed-effect model assumes that any variability in effects across studies is a function of sampling error. A random-effects model assumes two sources of variability in effects, one from sampling error and one for study level differences. More specifically, the random-effects model assumes a distribution of true population effects from which the observed studies are sampled (Hunter and Schmidt 2004; Lipsey and Wilson 2001a; Overton 1998; Raudenbush 1994). The random-effects model is recommended unless you have good reason to assume a fixed-effects model. Furthermore, a random-effects model will simplify to a fixed-effects model if the variability in effects across studies is homogeneous (Overton 1998).

The Mean Effect Size and Related Statistics

As with any statistical analysis, it is wise to examine a distribution for outliers or other anomalies. Outliers may represent genuinely unusual effects, often from small sample size studies, or they may reflect a computational error; as such, it is worth verifying the calculation of extreme effect-sizes. Both histograms and stem-and-leaf plots are useful graphical methods of examining the distribution of effect-sizes. Because analyses are weighted, an effect-size that represents an outlier may have a small effect on the overall analysis if it has a small relative weight (large variance). Of greater concern is an outlier with a moderate to large relative weight. I recommend performing sensitivity analyses that include and exclude any such effect-sizes.

The mean effect size under the fixed-effects model is simply a weighted mean, computed as

$$\overline{\mathrm{ES}} = \frac{\sum_{i=1}^{k} w_i \mathrm{ES}_i}{\sum_{i=1}^{k} w_i}, \tag{10.15}$$

where w is the inverse variance weight, and ES is the effect-size. The subscript i denotes the individual effect-sizes from 1 to k, where k is the number of effect-sizes.

Applying this equation to the data in Table 10.1 produces the following:

$$\overline{\mathrm{ES}} = \frac{55.956}{150.798} = 0.37.$$

Thus, the fixed-effects mean logged odds-ratio for these nine studies is 0.37. Taking the antilogarithm of this value converts the logged odds-ratio into a simple odds-ratio:

$$\mathrm{OR} = \mathrm{e}^{0.37} = 1.45.$$

The mean effect-size can be tested against the null hypothesis that the true population effect size equals zero using a z-test. To compute z, you need to compute the standard error of the mean effect-size:

$$\mathrm{se}_{\overline{\mathrm{ES}}} = \sqrt{\frac{1}{\sum_{i=1}^{k} w}}. \tag{10.16}$$

Recall that the inverse-variance weight for an individual effect-size is based on its standard error and a standard error is a statistical index of the precision of an effect-size. Thus, it is intuitive that the precision of the mean effect-size is a function of the precision of the effect-sizes on which it is based. The standard error of the mean effect-size (logged odds-ratio) for the data in Table 10.1 is

$$\mathrm{se}_{\overline{\mathrm{ES}}} = \sqrt{\frac{1}{150.798}} = 0.081.$$

Using the standard error of the mean effect-size, z is computed as

$$z = \frac{\overline{ES}}{\mathrm{se}}. \tag{10.17}$$

This tests the null hypothesis that the mean effect-size is zero. The mean effect-size is statistically significant at $p < 0.05$ if the z is greater than or equal to 1.96, assuming a two-tailed test. For a more precise p-value, consult the statistical tables in any good statistics book or use the built-in distribution functions of computer spreadsheets, mathematical programs, or statistical software packages. Applying this to the data in Table 10.1 produces

$$z = \frac{0.37}{0.081} = 4.57,$$

a z-value that is statistically significant at a conventional alpha level of 0.05. Under the assumptions of a fixed-effects model, we can clearly reject the null hypothesis that the population effect-size estimated by this collection of studies is zero.

Arguably, more informative than a significance test is a confidence interval, and in meta-analysis these are constructed in the usual manner using the standard error. The lower and upper bounds of a 95% confidence interval are computed as:

$$ES_{\text{lower}} = \overline{\text{ES}} - 1.96\text{se}, \tag{10.18}$$

$$ES_{\text{upper}} = \overline{\text{ES}} - 1.96\text{se}. \tag{10.19}$$

The 95% confidence interval for the mean effect-size for our working example is

$$\text{ES}_{\text{lower}} = \overline{0.37} - 1.96(0.081) = 0.21,$$

$$\text{ES}_{\text{upper}} = \overline{0.37} - 1.96(0.081) = 0.53.$$

We can be 95% confident that the true population effect estimated by these studies is between 0.21 and 0.53, under the assumptions of a fixed-effects model.

An important issue in meta-analysis is whether the distribution of effect-sizes is homogeneous or heterogeneous. A homogeneous distribution is one that varies no more than would be expected based on sampling error alone. In other words, the observed differences in the results across studies can be fully explained based on chance variation stemming from sampling error. A heterogeneous distribution is one that varies more than would be expected based on sampling error alone. This indicates that at least some portion of the observed differences in the effect-sizes across studies reflects true study effects. In statistical terms, there is a distribution of true population effects being estimated by these studies.

A statistical test of homogeneity is based on the chi-square distribution and is computed as

$$Q = \sum_{i=1}^{k} w_i \text{ES}_i^2 - \frac{\left(\sum_{i=1}^{k} w_i \text{ES}_i\right)^2}{\sum_{i=1}^{k} w_i}, \tag{10.20}$$

where the terms are defined as above. The degrees of freedom for Q is the number of effect-sizes (typically denoted as k) minus 1. The significance level is determined using the chi-square distribution. A statistically significant Q indicates that the distribution is *heterogeneous* and that the assumptions of the fixed-effects model are unjustified. The observed variability is greater than would be expected by chance suggesting that there are true differences across the studies.

Applying this equation to the effect-sizes in Table 10.1 is straightforward. The Q for these effect-sizes is

$$Q = 34.953 - \frac{55.956^2}{150.798} = 34.953 - 20.763 = 14.19,$$

with 8 degrees of freedom (the number of effect-sizes, k, minus 1). The p-value associated with the Q is $p = 0.077$; as such, it is not statistically significant at a conventional alpha level. We cannot reject the null hypothesis that these effect-sizes are homogeneous.

A weakness of the Q test is that it is statistically underpowered in cases where the number of studies is small, such as in this example. This often results in a failure to identify true heterogeneity. Higgins et al. (2003) have proposed an alternative index of heterogeneity, I^2, recommended for use when the number of studies is small. This index is computed as

$$I^2 = 100\% \times \frac{Q - df}{Q} \tag{10.21}$$

and ranges from 0 to 100%. If Q is less than df, then I^2 is set to 0%. The larger the value of I^2, the more heterogeneity, with the values of 25%, 50%, and 75% roughly representing low, moderate, and high levels of heterogeneity. Applying this to the effect-sizes in Table 10.1 suggests that this distribution has moderate heterogeneity:

$$I^2 = 100\% \times \frac{14.19 - 8}{14.19} = 44\%. \tag{10.22}$$

Computing the above statistics under a *random-effects* model involves first estimating the random-effects (or between study) variance component (τ^2), of which there are several estimators (see Schulze 2004; Raudenbush 2009; Viechtbauer 2005). The most commonly used one was developed by DerSimonian and Laird (1986) and is a closed-form method-of-moments estimator. Not surprisingly, it is based on the value Q, the estimate of heterogeneity. The DerSimonian and Laird formula for the random-effects variance component is

$$\tau^2 = \frac{Q - (k-1)}{\sum w - \frac{\sum w^2}{\sum w}}, \tag{10.23}$$

where the terms are defined as above. Notice that when Q is less than $k - 1$, τ^2 becomes negative. Because it is not possible to have negative variability, negative values of τ^2 are set to zero. The expected value for a chi-square is its degrees of freedom. Hence, if Q is greater than its degrees of freedom ($k - 1$), then there is greater variability than expected under the null hypothesis, even if not statistically significantly so. A Q that is less than $k - 1$ indicates that there is less variability than expected under the null hypothesis. In this way, τ^2 reflects the excess variability in the effect-size distribution.

Under the fixed-effects model, the inverse-variance weight only reflects variability due to sampling error. The random-effects model assumes that the precision of an effect-size is a function not only of sampling error but also true variability in effects across studies in the population from which these studies were drawn. Thus, we must add this variability to the estimated sampling error variability and recompute the inverse-variance weight. The random-effects inverse-variance weight for each effect-size is defined as,

$$w_i = \frac{1}{v_i + \tau^2}, \tag{10.24}$$

where v is the appropriate formula above (i.e., (10.10), (10.12), (10.13), or (10.14)). The mean effect-size, z-test, and confidence intervals are computed using these new weights. The homogeneity test, Q, is not computed with these new weights as the excess variability is now incorporated into the model.

Most of the values needed to compute τ^2 for the data in Table 10.1 have already been computed with the exception of the sum of the squared weights. Squaring the weights and summing produces the value 63, 848.51. Using this value and those determined previously, τ^2 equals

$$\tau^2 = \frac{14.19 - (9-1)}{150.798 - \dfrac{5,899.210}{150.798}} = \frac{6.19}{111.669} = 0.0554.$$

This value is added to the variance estimate for each effect-size and a new inverse-variance weight is computed. Doing so for the first row in Table 10.1 produces

$$w_i = \frac{1}{0.579 + 0.0554} = 1.576. \tag{10.25}$$

For our working example, the random-effects mean is 0.40 with a 95% confidence interval of 0.14 to 0.66. The random-effects mean is slightly higher than the fixed-effects mean of 0.37 and the confidence interval is larger. The latter reflects the more conservative nature of a random-effects model when compared with a fixed-effects model. There is increased uncertainty incorporated into the model based on the variability in effect-sizes across studies.

The DerSimonian and Laird (1986) method-of-moments estimator for τ^2 is considered unbiased, but is less efficient than other estimators, such as the restricted maximum likelihood estimator. However, as discussed in Friedman (2000), Shadish and Haddock (2009), and Viechtbauer (2005), the DerSimonian and Laird method is suitable for many typical situations in meta-analysis, except when the study sample sizes are large or the heterogeneity between studies is large (greater than 3 times the degrees of freedom of Q). The maximum likelihood estimator is more efficient but negatively biased when the number of studies is small. Viechtbauer (2005) argued that the restricted maximum likelihood estimator strikes a good balance between efficiency and bias and I would recommend it, at least for situations where the DerSimonian and Laird estimator is known to be weak. Given the iterative nature of the restricted maximum likelihood estimator, I recommend relying on computer software implementations for performing analyses based on these alternative estimators of τ^2.

Moderator Analyses

Rarely is a meta-analysis focused solely on the overall mean effect-size. Frequently, one is also interested in examining the relationship between study features of the variability in results across studies. For example, you may be interested in examining whether characteristics of the studies relate to the observed effects sizes, or in a meta-analysis of a specific offender treatment program, you may be interested in examining whether some program elements are more effective than others. This is the focus of moderator analysis and there are two main analytic approaches. The first compares the means across a categorical grouping of studies, such as treatment type. The second relies on multiple regression methods to examine either a continuous moderator variable, such as program size or year of publication, or multiple moderator variables.

A Single Categorical Moderator. Studies included in a meta-analysis can often be categorized into meaningful subgroups either on a substantive or a methodological variable, such as treatment type, restriction of the sample to high-risk offenders, etc. These categorical variables can serve as the independent variable in a moderator analysis that is analogous to a t-test when the independent variable has only two categories and a one-way analysis-of-variance (ANOVA) when the independent variable has three or more categories. Of interest is whether the means differ across the categories. Hedges and Olkin (1985) developed an analog-to-the-ANOVA that tests this hypothesis.

The analog-to-the-ANOVA partitions the total variability in effect-sizes into two portions: that between the groups and that within the groups. The Q of (10.20) represents the total variability in effect-sizes, that is, the variability of effect-sizes around the overall mean effect-size. The variability within each groups is the variability of effect-sizes around the group means. Hence, the Q_{within} is computed as

$$Q_{\text{within}} = \sum w_{ij} \left(\text{ES}_{ij} - \overline{\text{ES}_j}\right)^2, \tag{10.26}$$

where j denotes the individual groups or categories of the independent variable. Essentially, this computes a Q within each group and then sums those Qs across groups. The Q_{between} can be computed through subtraction as

$$Q_{\text{between}} = Q_{\text{total}} - Q_{\text{within}} \tag{10.27}$$

where Q_{total} is the overall Q from (10.20). Q_{between} can also be computed directly as

$$Q_{\text{between}} = \sum_{j=1}^{p} \left(\overline{\text{ES}}_j w_j\right)^2 - \frac{\left(\sum_{j=1}^{p} \overline{\text{ES}}_j w_j\right)}{\sum_{j=1}^{p} w_j}, \tag{10.28}$$

where $\overline{\text{ES}}_j$ is the mean effect-sizes for each group, and w_j is the sum of the weights within each group. The degrees of freedom for Q_{between} is the number of groups (p) minus 1, just as with the F_{between} in a one-way ANOVA. Similarly, the degrees of freedom for the Q_{within} is the number of studies (k) minus the number of groups (p). The degrees of freedom between and within should sum to the degrees of freedom total, or $k - 1$. As with Q_{total}, Q_{between} and Q_{within} are distributed as chi-squares.

A statistically significant Q_{between} is interpreted in the same way as an F from a one-way ANOVA or as an independent t-test if the independent variable has only two categories. In the latter case, a significant Q_{between} indicates that the difference between the two mean effect-sizes is statistically significant. For an independent variable with three or more groups, a significant Q_{between} indicates that the independent variable explains significant variability across the effect-sizes, or more simply that the mean effect-sizes differ across groups. Focused contrasts between means, such as pairwise comparisons, can be run to identify the source of the variability, just as with a one-way ANOVA.

Under a fixed-effects model, the Q_{within} is also meaningful, as it indicates whether the residual variability in effect-sizes remains heterogeneous after accounting for the independent variable. A statistically nonsignificant Q_{within} indicates that the moderator variable reduced the variability across effect-sizes to not more than what would be expected because of chance variation. A statistically significant Q_{within} indicates that the variability in the effect-sizes within the groups remains heterogeneous. In which case, you should consider fitting a random-effects version of the analog-to-the-ANOVA.

A random-effects analog-to-the-ANOVA is sometimes referred to as a *mixed-effects* model. This is because the categorical or independent variable is treated as fixed and the variability within the groups is treated as random. In terms of a general linear model, this would be a fixed-slopes, random-intercept model. As with the overall mean, the mixed-effects model is based on a recomputed inverse-variance weight that includes a random-effects variance component representing the variability across studies. In the case of moderator analysis, however, only the variability across studies which is *not* explained by the independent variable is used. Thus, the method-of-moments estimator for the random-effects variance component (τ^2) is based on Q_{within} rather than from Q_{total}. The denominator for τ^2 becomes more complicated and is not presented here (see Raudenbush 1994). The computer macros discussed below for SPSS, Stata, and SAS, as well as meta-analysis computer programs such as comprehensive meta-analysis (http:\www.meta-analysis.com) implement several of these methods. With the new set of inverse-variance weights, the analog-to-the-ANOVA model is refit, producing new values for Q_{between}, Q_{within}, and the mean effect-size and related statistics within each group. It is important to recognize that the Q_{within} produced using the random-effects weights is not meaningful, as the excess variability is now incorporated into the weights. The Q_{within} from the fixed-effects model is the correct test for whether the moderator variable accounts for the excess variability in the distribution of effect-sizes.

Figure 10.1 shows the output from a Stata macro called *metaf* available at http://mason.gmu.edu/~dwilsonb/ma.html. This moderator analysis compared the mean effect-size from studies with a randomized or true experimental design with the mean effect-size from studies with a nonrandomized or quasiexperimental design. The output shows that the test of the difference between the means was not statistically significant ($Q_{\text{between}} = 0.007$, df $= 1$, $p = 0.93$). The mean effect-size for the two design types was roughly equal (0.43 vs. 0.40, respectively). Thus, design type was not related to the observed effect-size.

```
. metaf lgor random [w=wlgor], model(mm) Version 2005.05.23 of
metaf.ado

Meta-Analytic Analog to the One-way ANOVA, Mixed Effects Model
---------------------------------------------
  Source |            Q        df          P
---------------------------------------------
 Between |       0.0070         1    0.93319
  Within |       8.6461         7    0.27907
---------------------------------------------
   Total |       8.6531         8    0.37240

Descriptive Fixed Effects Meta-Analytic Results by: random
-------------------------------------------------------------------------
random    |      Mean  St. Er.  [95% Conf. Int.]        z    P>|z|        k
-------------------------------------------------------------------------
0         |    .42787   .26154  -.08475   .94048   1.6359  0.10185
4 1         |    .40181   .16799   .07255   .73107   2.3918
0.01677        5
-------------------------------------------------------------------------
Total     |    .40942   .14135   .13238   .68645   2.8965  0.00377
9

Mixed Effects Homogeneity Analysis by: random
---------------------------------------------
  Source |           Qw        df          P
---------------------------------------------
0        |       0.2890         3    0.98274 1         |
8.3571         4    0.12107
---------------------------------------------
Random effects variance component (via method of moments) =
.0716085
```

FIGURE 10.1. Output from Stata showing an analog-to-the-ANOVA type moderator analysis of the relationship between design type (Random = 1; Nonrandom = 0) and effect size.

A Continuous Moderator or Multiple Moderators. Continuous variables and multiple moderator variables fit naturally within a multiple regression framework. For example, it may be meaningful to examine whether the number of sessions of a cognitive-behavioral program is related to the observed reduction in reoffending or whether the year of data collection is related to the effect of race on an officer's decision to make an arrest. Additionally, it may be meaningful to examine multiple moderator variables in a single analysis given the often highly confounded nature of study characteristics. Unfortunately, standard OLS regression is not appropriate for meta-analytic data, as it produces incorrect standard errors and associated inferential statistics (Hedges and Olkin 1985). Under a fixed-effects model, it is possible to adjust the results of a standard weighted least squares regression model (see Hedges and Olkin 1985; Lipsey and Wilson 2001a). However, I recommend that you simply use the computer macros discussed in this chapter or a meta-analysis computer program that performs meta-analytic regression.

Meta-analytic regression, or meta-regression as some call it, is specifically designed for meta-analytic data. Both fixed- and random-effects regression models can be estimated. As with the analog-to-the ANOVA, two Q values are estimated, one for the model and one for the residuals. The Q_{model} is interpreted in the same fashion as the overall F from OLS regression. A statistically significant Q_{model} indicates that the linear combination of moderator variables explains significant variability in the observed effects. Under the fixed-effects model, a statistically significant Q for the residuals indicates that the residual distribution in effect-sizes remains heterogeneous, suggesting that the assumptions of a fixed-effects model are likely false and a random-effects (mixed-effects) model should be used.

Figure 10.2 shows the results of a regression analysis using data from Table 10.1. The two independent variables are a dummy code for treatment type (Moral Reconation, coded as 0, and Reasoning and Rehabilitation, coded as 1) and a dummy code reflecting whether the treatment occurred within a prison/jail or a community setting (1 = yes; 0 = no). The results show that these variables do not significantly predict the observed effect-size, either as an overall model ($Q_{\text{model}} = 1.4158$, df $= 1$, $p = 0.49$), or individually (both regression coefficients are statistically nonsignificant at a conventional level). It is worth noting, however, that the regression coefficients are of a meaningful size. The coefficient for treatment type

```
. metareg lgor txtype institution [w=wlgor], model(mm)
Version 2005.05.23 of metareg.ado

Meta-Analytic Random Intercept, Fixed Slopes Regression Analysis

  Source |            Q        df          P            No. of obs   =       9
------------------------------------------                  Mean ES = 0.4070
   Model |       1.4158         2    0.49267              R-squared = 0.1605
Residual |       7.4064         6    0.28489
------------------------------------------
   Total |       8.8222         8    0.35752

----------------------------------------------------------------------------
Variable |          B          SE          z      P>|z|   [95% Conf. Interval]
----------------------------------------------------------------------------
txtype     |    -.39058     .333902   -1.16973   0.242108   -1.04502    .263872
institution|  -.28644     .345596    -.82884   0.407195   -.963813    .390925
_cons      |     .80564     .383221    2.10228   0.035529    .054526    1.55675
----------------------------------------------------------------------------
Random effects variance component (via method of moments) =  .0661342
```

Figure 10.2. Output from Stata showing a multiple regression type moderator analysis regressing effect size on the type of program (Moral Reconation vs. Reasoning and Rehabilitation) and whether the program was in a prison or jail.

predicts a −0.391 difference in the logged odds-ratio between the treatment types, with Moral Reconation producing the larger effect. The five programs that were conducted in a prison or jail produced effects that are roughly −0.286 smaller than four programs conducted in a community setting. The small number of effect-sizes resulted in low statistical power for the moderator analysis. Clearly, these differences are worthy of future research but must be interpreted with caution. We cannot rule out sampling error as an explanation. Furthermore, with moderator analysis, it is always possible that some other feature of the studies is confounded with the moderator variable and could be responsible for the difference in effects (see Lipsey and Wilson 2001b).

Publication Bias

Publication bias is recognized as a particularly significant threat to the validity of meta-analysis (Wilson 2009). This bias stems from the well documented phenomenon that statistically significant results are more likely to be published than nonsignificant results (see Cooper 1998; Dickersin 2005; Gerber and Malhotra 2008; Lipsey and Wilson 2001a, 1993; Rosenthal 1991; Wilson 2009). There are several mechanisms for this effect. First, authors may be more likely to write-up and submit for publication results that are statistically significant. Second, peer reviewers and editors tend to be more enthusiastic about publishing statistically significant results than nonsignificant results. Third, statistically significant results are more likely to be cited by other studies making them more easily identified as part of the review process. Publication-selection bias not only affects whether an entire study is included in a meta-analysis but which outcomes and analyses are included. Primary authors are more likely to report those outcomes or statistical models that produced significant results than those that did not. There is substantial empirical evidence that publication-selection bias occurs throughout the social and medical sciences (see Dickersin 2005; Lipsey and Wilson 1993; Stern and Simes 1997). This problem affects all methods of taking stock of the research literature, whether it is through a traditional narrative review or a meta-analysis.

The first line of defence against publication-selection bias is to search for and include both published and unpublished studies that meet explicit criteria. Unpublished sources might include dissertations, technical reports, government reports, and conference presentations. A common concern that I hear when advocating for the inclusion of unpublished studies in a meta-analysis is that these studies may be of inferior quality. First, this presumes that unpublished studies are studies that were rejected by the peer review process because of methodological flaws. Many unpublished studies were never submitted to an academic journal and thus never rejected based on methodological flaws (Cooper et al. 1997). Second, much good work in criminology and criminal justice is conducted in research organizations where publishing in peer reviewed journals is less of a priority. Once a technical report is submitted to the funding agency, the authors may be too busy working on new projects to pursue publication. Finally, the peer review process does not have unique information about the methodological quality of a study. Assessments of methodological quality are based on the written manuscript as part of the peer review process and as part of a meta-analysis. The criteria for determining which studies to include and exclude should specify the necessary methodological characteristics, including any methodological flaws that would exclude a study from consideration. The nature of the research question should determine how strict or lenient these criteria are.

Assuming you have included unpublished studies in your meta-analysis, you should still examine the data for evidence of publication-selection bias. There are several ways to do this. First, you can simply examine the mean effect-size for published and unpublished studies. A large difference suggests that publication bias is likely to be present in your data. For example, four of the studies reported in Table 10.1 were peer reviewed journal articles and five were from unpublished works (two), government reports (one), and book chapters (two). The mean effect-size for the journal articles was 0.62, whereas it was 0.19 for the latter publication types, suggesting the presence of publication-selection bias. Had we only included published studies, we would have overestimated the effectiveness of these programs.

A second method is the funnel plot (Sterne et al. 2005). This is a simple scatter plot of the effect-size against the standard error of the effect-size ($\mathrm{se}_{\mathrm{ES}} = \sqrt{v_{\mathrm{ES}}}$ and $v_{\mathrm{ES}} = 1/w$). Generally, the effect-size is on the x-axis and the standard-error of the effect-size is on the y-axis. The scatter plot should look like a funnel. The logic of this is that the effect-sizes with small standard errors (large sample size studies) should vary little around the overall mean effect-size whereas study with large standard errors (small sample size studies) should vary substantially around the overall mean effect-size. Publication bias is evident when this scatter plot is asymmetric around the mean effect-size. In particular, publication bias is most evident when effects for larger standard error studies are missing near the null value (i.e., effects that would not have been statistically significant).

A third method proposed by Egger et al. (1997) and Sterne et al. (2005) examines the linear relationship between the standard normal deviate of the effect-size and its precision. The standard normal deviate is the effect-size divided by its standard error

$$z_i = \frac{\mathrm{ES}_i}{\mathrm{se}_i}. \tag{10.29}$$

Precision is the inverse of the standard error

$$\mathrm{prec}_i = \frac{1}{\mathrm{se}_i}. \tag{10.30}$$

Thus, the smaller the precision of an effect-size, the smaller the value of $prec_i$. The standard normal deviate of the effect-size is then regressed on the precision of the effect-size using OLS regression. A negative relationship between effect-size and precision is suggestive of publication bias. The logic of this method is that smaller effects will be statistically significant in larger studies and as such more likely to be observed. Unfortunately, a weakness of this method is that there are other reasons why effect-size might be related to precision. Larger studies often differ from smaller studies in important ways that might affect the observed effect. In program effectiveness research, larger studies may have more problems with the integrity of the intervention or have a more diverse and less ideally selected group of participants than a smaller study. Thus, a relationship between effect-size and precision is evidence of publication-selection bias, but there are often other plausible explanations for the relationship.

Forest-Plots

A powerful graphical representation of meta-analytic data is the forest-plot. The forest-plot presents the effect-size and 95% confidence interval for each study and typically also

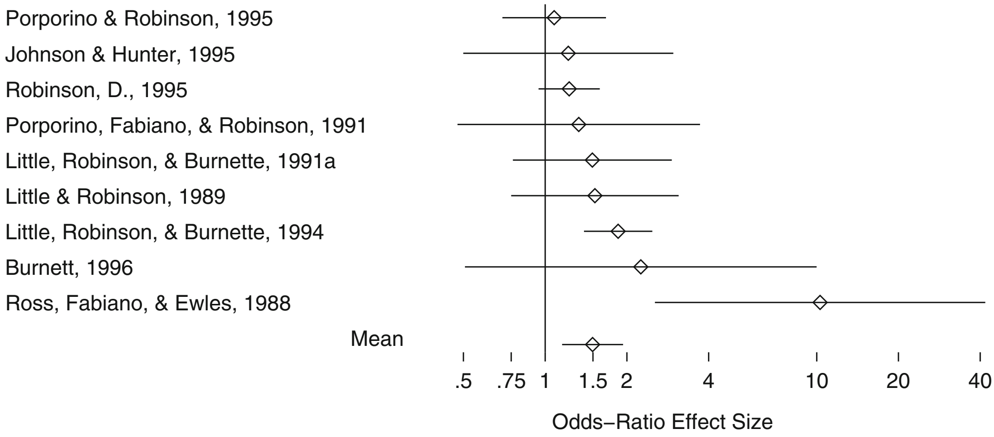

FIGURE 10.3. Forest-plot of odds-ratios and 95% confidence intervals for the effects of cognitive-behavioral programs on recidivism.

includes the overall mean effect-size and confidence interval. A forest-plot of the group-based cognitive-behavioral programs for offenders is shown in Fig. 10.3. The diamonds in this figure show the value of the odds-ratio effect-size for each study and the horizontal line through each diamond represents the 95% confidence interval. The studies are sorted by the size of the odds-ratio. We can clearly see that all of the studies observed a positive odds-ratio (i.e., odds-ratio greater than 1) and the pattern of results suggests that these programs are effective overall. The mean odds-ratio is shown at the bottom and is roughly 1.5 with a 95% confidence interval that does not include the null value of 1. As such, it is statistically significant at $p < 0.05$ and suggests a reduction in the odds of reoffending of roughly 50% attributable to the cognitive-behavioral programs.

Figure 10.3 is a "bare-bones" forest-plot. More complex forest-plots can include the sample size of the studies or the inverse-variance weights. Studies may be grouped by an important moderator variable with mean effect-sizes reported for each grouping. Another common variation is to size the diamond as a function of the inverse-variance weight. This helps visually depict the weight each effect has on the overall mean. Regardless of its complexity, forest-plots are excellent tools for communicating the results of meta-analysis to a broad audience.

Multivariate Research

The meta-analysis of multivariate research presents unique challenges. The methods discussed earlier assume a single statistical parameter of interest, based either on a single variable, such as a rate, or a bivariate relationship between two variables, as is represented in a correlation or odds-ratio. Many research questions in criminology involve multivariate statistical models and it is desirable to be able to meta-analyze collections of such studies. I will distinguish between two types of multivariate research, as the meta-analytic issues differ between them. First are basic multiple regression designs with a single dependent variable and two or more independent variables where the research focus is on one independent variable and the additional independent variables are included in the model to control for spuriousness. Second are multivariate statistical models with multiple dependent variables and independent variables, such as a structural equation model.

Multiple regression and related methods, such as logistic regression, are widely used in criminology and the social sciences more generally. This statistical method tests the effect of multiple independent variables on a dependent variable. The regression coefficient for each independent variable represents a bivariate effect-size between a specific independent variable and the dependent variable, adjusted for any confounding with the other independent variables. The standardized regression coefficient, β, from an ordinary least squares (OLS) regression model is a usable effect-size and can be interpreted as an r, adjusted for the other independent variables in the model.

There are, however, several complications in meta-analyzing standardized regression coefficients from OLS models. First, each study is likely to include a different set of independent variables. This will produce heterogeneity in the βs across studies. As such, an important theoretical question is whether these variations in the models across studies produce effect-sizes that are incommensurate. It may be possible to treat the variation in models as a moderator in the meta-analysis, depending on the amount of variation and the number of studies available for analysis. Second, not all studies will report the standardized regression coefficient, opting instead to report the unstandardized form. Given the right descriptive statistics, β can be computed as

$$\beta = B\frac{s_x}{s_y}, \tag{10.31}$$

where B is the unstandardized regression coefficient, s_x is the standard deviation for the independent variable, and s_y is the standard deviation for the dependent variable. Third, the variance or standard error for β is often not provided. The standard error of the unstandardized regression coefficient is often provided and this can be converted to the standard error of β with (10.31). Recall that the variance is simply the standard error squared. The inverse-variance weight for z (10.10) can be used but will overestimate the variance and underweight the effect-size. A better approximation is

$$\mathrm{se}_\beta = \sqrt{\frac{1 - R_y^2}{n - p - 1}}, \tag{10.32}$$

where R_y^2 is the multiple R-squared for the regression model, n is the sample size, and p is the number of predictors. This will still overestimate the standard error but to a lesser degree than (10.10). The overestimation is a function of the degree of covariation with the other variables in the regression model, as shown by

$$\sqrt{\frac{1}{1 - R_x^2}}, \tag{10.33}$$

where R_x^2 is the variance in the independent variable of interest explained by all other independent variables in the model (Cohen and Cohen 1983).

Meta-analyzing logistic regression models is less complicated. As with OLS regression, the models are likely to include different predictors across studies, adding heterogeneity to the effect-size. However, the unstandardized regression coefficient is a logged odds-ratio and can be used directly as the effect-size. Studies reporting on the results from a logistic regression model typically provide both the unstandardized regression coefficient and its associated standard error.

Meta-analyzing truly multivariate research involving multiple dependent variables and both direct and indirect effects requires an alternative model-based approach (Becker 2009). The fundamental goal is to estimate a common correlation matrix across studies and use this synthesized correlation matrix as the basis for multivariate analysis, whether that is factor analysis, structural equation modeling, or regression analysis. This model-based approach has also been called meta-analytic structural equation modeling (Cheung and Chan 2005; Furlow and Beretvas 2005) and two-stage structural equation modeling (Cheung and Chan 2005). A restricted application of this approach includes only studies that measure all of the variables of interest. In more common usage, most studies will only estimate portions of the correlation matrix with few or no studies estimating all of the correlations. There are numerous statistical challenges in producing a synthesized correlation matrix and several methods have been developed for conducting these model-based meta-analyses (see Becker 2009; Cheung and Chan 2005). This is an area of meta-analysis with great potential given the ubiquity of multivariate models in the social sciences.

COMPUTER SOFTWARE

There are several computer programs designed specifically for meta-analysis. I am most familiar with Comprehensive Meta-Analysis[2] (http://www.meta-analysis.com). This program can perform all of the analyses discussed in this chapter and can also produce publication quality forest-plots and funnel-plots. Other similar programs include RevMan developed by the Cochrane Collaboration (http://www.cochrane.org), Meta-Analysis (Version 5.3) developed by Ralf Schwarzer (http://userpage.fu-berlin.de/~health/meta_e.htm), Meta-Win (http://metawinsoft.com), WEasyMa (http://www.weasyma.com), and DSTAT (http://www.erlbaum.com). This is by no means an exhaustive list.

An alternative to using a free-standing software program is to use macros to extend the capabilities of existing statistical software packages. For example, I have developed a set of macros for SAS, SPSS, and Stata that perform the analyses discussed in this chapter. See http://mason.gmu.edu/~dwilsonb/ma.html for more details. These macros have the advantage that you can make use of all the data manipulation capabilities of the main software program. This flexibility is useful when conducting a large meta-analysis with possibly numerous effect-sizes per study. Using standard data manipulation procedures, subsets of statistically independent effect-sizes can be generated and meta-analyzed.

CONCLUSION

Since the seminal pieces of research on meta-analysis in the mid and late 1970s (see Hunt (1997) for a discussion of the development of meta-analysis), meta-analysis has become an increasingly popular method of synthesizing the empirical evidence on a research question

[2] I have served as a consulted in the development of this program.

throughout the social sciences, medicine, public health, ecology, and other fields. I believe that this growth stems from the many advantages of meta-analysis over more traditional review methods.

First, the systematic review methods of meta-analysis allow for a transparency and replicability that are valued in the sciences. A properly executed meta-analysis is explicit about the study inclusion and exclusion criteria, the methods for searching for studies, the coding process, and the statistical methods. This allows for critical scrutiny by other scholars and replication. Author judgments are still a critical part of the process, such as defining the boundaries for the review, or deciding which moderator analyses to explore. However, these decisions are explicit and open to scholarly debate.

Second, the focus on the direction and magnitude of the effects across studies, rather than simply obsessing over statistical significance, provides more detailed information about the empirical findings of interest and does so in a way that allows for comparisons across studies. Statistical significance confounds sample size and the size of the empirical relationship of interest. The former is not a feature of the phenomenon under study but something under the control of the researcher. The use of effect-size appropriately shifts the focus to the direction and magnitude of effect, which is, after all, what is of greatest interest.

Third, the methods of meta-analysis allow for the exploration of variability in results across studies. In its simplest form, this illuminates whether findings are consistent or not across studies. Variation in effects across studies can be explored through moderator analyses. These analyses can assess whether substantive or methodological differences across studies explain variation in observed effects. The statistical methods also help protect against interpretation of differences across studies that may simply reflect chance variation.

Fourth, the detailed coding and comprehensive search methods that are integral to a well done meta-analysis allow for the assessment of gaps in the existing knowledge base. By carefully cataloging what is known, meta-analysis provides a firm foundation for identifying areas of needed research. A meta-analyst may also conclude that a specific research question has been adequately answered by the extant research base and that future research should focus on extending the knowledge-base in new ways.

Fifth, meta-analysis can handle a large number of studies. Although meta-analysis can be productively applied to a small number of studies (i.e., 2–10 studies), it is robust enough to handle hundreds of studies. The larger the meta-analysis, the greater the opportunity for complex moderator analyses.

The methods of meta-analysis have greatly advanced over the past 30 years and will continue to be refined and extended. An exciting area of potential advancement is in the handling of multivariate research questions. Although there is debate about the validity of specific statistical methods in meta-analysis (see, for example Berk 2007; Lipsey 2007; Shadish 2007), this approach provides a superior alternative to traditional narrative review methods that place too much emphasis on statistical significance and unusual findings that may simply be statistical outliers. The appeal of meta-analysis is spreading to disciplines beyond social science and medicine, such as ecology. I expect this trend to continue.

REFERENCES

Becker BJ (2009) Model-based meta-analysis, 2nd edn, Russell Sage Foundation, New York

Berk R (2007) Statistical inference and meta-analysis. J Exp Criminol 3:247–270, doi10.1007/s11292-007-9036-y, URL http://dx.doi.org/10.1007/s11292-007-9036-y

Borenstein M (2009) Effect sizes for studies with continuous outcome data, 2nd edn. Russell Sage Foundation, New York

Burnett WL (1996) Treating postincarcerated offenders with Moral Reconation Therapy^{TM}: A one-year recidivism study. PhD thesis, University of Phoenix

Bushway SD, Sweeten G, Wilson DB (2006) Size matters: standard errors in the application of null hypothesis significance testing in criminology and criminal justice. J Exp Criminol 2:1–22, doi: 10.1007/s11292-005-5129-7, URL http://dx.doi.org/10.1007/s11292-005-5129-7

Cheung MWL, Chan W (2005) Meta-analytic structural equation modeling: A two-stage approach. Psychol Methods 10:40–64. doi: 10.1037/1082-989X.10.1.40, URL http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=met-10-1-40&site=ehost-live

Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral sciences, 2nd edn. L. Erlbaum Associates, Hillsdale, NJ

Cooper HM (1998) Synthesizing research: A guide for literature reviews. Sage, Thousand Oaks, CA

Cooper HM, DeNeve K, Charlton K (1997) Finding the missing science: The fate of studies submitted for review by a human subjects committee. Psychol Methods 2:447–452. doi: 10.1037/1082-989X.2.4.447, URL http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=met-2-4-447&site=ehost-live

Cox DR (1970) The analysis of binary data. Methuen, London

DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 7:177–88. doi: 3802833, URL http://www.ncbi.nlm.nih.gov/pubmed/3802833

Dickersin K (2005) Publication bias: recognizing the problem, understanding its origins, and preventing harm. Wiley, Chichester, pp 11–33

Egger M, Smith GD, Schneider M, Minder C (1997) Bias in meta-analysis detected by a simple, graphical test. BMJ 315:629–634, URL http://www.bmj.com/cgi/content/abstract/315/7109/629

Field AP (2001) Meta-analysis of correlation coefficients: A monte carlo comparison of fixed- and random-effects methods. Psychol Methods 6:161–180. doi: 10.1037/1082-989X.6.2.161, URL http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=met-6-2-161&site=ehost-live

Fleiss JL (1994) Measures of effect size for categorical data. Russell Sage Foundation, New York, pp 245–260

Fleiss JL, Berlin JA (2009) Measures of effect size for categorical data, 2nd edn. Russell Sage Foundation, New York

Friedman L (2000) Estimators of random effects variance components in meta-analysis. J Educ Behav Stat 25:1–12. doi: 10.3102/10769986025001001, URL http://jeb.sagepub.com/cgi/content/abstract/25/1/1

Frost JJ, Forrest JD (1995) Understanding the impact of effective teenage pregnancy prevention programs. Fam Plann Perspect 27:188–195, URL http://mutex.gmu.edu:2112/stable/2136274

Furlow CF, Beretvas SN (2005) Meta-analytic methods of pooling correlation matrices for structural equation modeling under different patterns of missing data. Psychol Methods 10:227–254, doi: 10.1037/1082-989X.10.2.227, URL http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=met-10-2-227&site=ehost-live

Gerber AS, Malhotra N (2008) Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? Sociol Methods Res 37:3–30, doi: 10.1177/0049124108318973, URL http://smr.sagepub.com/cgi/content/abstract/37/1/3

Glass GV (1976) Primary, secondary, and meta-analysis research. Educ Res 5:3–8

Gleser LJ, Olkin I (1994) Stochastically dependent effect sizes. Russell Sage Foundation, New York, pp 339–356

Hasselblad V, Hedges (1995) Meta-analysis of screening and diagnostic tests. Psychol Bull 117:167–178. doi: 10.1037/0033-2909.117.1.167, URL http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=bul-117-1-167&site=ehost-live

Hedges LV (1981) Distribution theory for Glass's estimator of effect size and related estimators. J Educ Behav Stat 6:107–128, doi: 10.3102/10769986006002107, URL http://jeb.sagepub.com/cgi/content/abstract/6/2/107

Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic, Orlando, FL

Higgins JPT, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327:557–560. doi: 10.1136/bmj.327.7414.557, URL http://www.bmj.com

Hunt MM (1997) How science takes stock: The story of meta-analysis. Russell Sage Foundation, New York

Hunter JE, Schmidt FL (1990) Methods of meta-analysis: Correcting error and bias in research findings. Sage Publications, Newbury Park

Hunter JE, Schmidt FL (2004) Methods of meta-analysis: Correcting error and bias in research findings, 2nd edn. Sage, Thousand Oaks, CA

Johnson G, Hunter RM (1995) Evaluation of the specialized drug offender program. In: Ross RR, Ross B (eds) Thinking straight. Cognitive Center, Ottawa, ON, pp 215–234

Kalaian HA, Raudenbush SW (1996) A multivariate mixed linear model for meta-analysis. Psychol Methods 1:227–235. doi: 10.1037/1082-989X.1.3.227, URL http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=met-1-3-227&site=ehost-live

Kuhns JB, Wilson DB, Maguire ER, Ainsworth SA, Clodfelter TA (2008) A meta-analysis of marijuana, cocaine, and opiate toxicology study findings among homicide victims. Addiction 104:1122–1131

Lipsey MW (1995) What do we learn from 400 research studies on the effectiveness of treatment with juvenile delinquents? Wiley, New York, pp 63–78

Lipsey MW (2007) Unjustified inferences about meta-analysis. J Exp Criminol 3:271–279. doi: 10.1007/s11292-007-9037-x, URL http://dx.doi.org/10.1007/s11292-007-9037-x

Lipsey MW, Cullen FT (2007) The effectiveness of correctional rehabilitation: A review of systematic reviews. Annu Rev Law Soc Sci 3:297–320, doi: 10.1146/annurev.lawsocsci.3.081806.112833, URL http://mutex.gmu.edu:2078/doi/full/10.1146/annurev.lawsocsci.3.081806.112833

Lipsey MW, Derzon JH (1998) Predictors of violent or serious delinquency in adolescence and early adulthood: A synthesis of longitudinal research. Sage Publications, Thousand Oaks, CA, pp 86–105

Lipsey MW, Wilson DB (1993) The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. Am Psychol 48:1181–1209. doi: 10.1037/0003-066X.48.12.1181, URL http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=amp-48-12-1181&site=ehost-live

Lipsey MW, Wilson DB (1998) Effective intervention for serious juvenile offenders: A synthesis of research. Sage Publications, Thousand Oaks, CA, pp 313–345

Lipsey MW, Wilson DB (2001a) Practical meta-analysis. Applied social research methods series. Sage Publications, Thousand Oaks, CA

Lipsey MW, Wilson DB (2001b) The way in which intervention studies have "personality" and why it is important to meta-analysis. Eval Health Prof 24:236–254. doi: 10.1177/016327870102400302, URL http://ehp.sagepub.com/cgi/content/abstract/24/3/236

Lipsey MW, Crosse S, Dunkle J, Pollard J, Stobart G (1985) Evaluation: The state of the art and the sorry state of the science. New Dir Program Eval 27:7–28

Little GL and Robinson KD (1989) Treating drunk drivers with Moral Reconation therapy: A one-year recidivism report. Psychol Rep 64:960-962

Little GL, Robinson KD Burnette KD (1991) Treating drug offenders with Moral Reconation therapy: A three-year recidivism report. Psychol Rep 69:1151–1154

Little GL, Robinson KD, Burnette KD (1994) Treating offenders with cognitive-behavioral therapy: 5-year recidivism outcome data on MRT. Cogn Behav Treat Rev 3:1-3

Overton RC (1998) A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. Psychol Methods 3:354–379. doi: 10.1037/1082-989X.3.3.354, URL http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=met-3-3-354&site=ehost-live

Porporino FJ, Robinson D (1995) An evaluation of the Reasoning and Rehabilitation program with Canadian federal offenders. In: Ross RR, Ross B (eds) Thinking straight. Cognitive Centre, Ottawa, ON, pp 155–191

Porporino FJ , Fabiano EA, Robinson D (1991) Focusing on successful reintegration: Cognitive skills training for offenders, r19. Ottawa, Canada: Research and Statistics Branch, The Correctional Service of Canada

Pratt TC, Cullen FT (2000) The empirical status of Gottfredson and Hirschi's General Theory of crime: A meta-analysis. Criminology 38:931–964, doi: 10.1111/j.1745-9125.2000.tb00911.x, URL http://mutex.gmu.edu:2167/doi/abs/10.1111/j.1745-9125.2000.tb00911.x

Raudenbush SW (1994) Random effects models. Russell Sage Foundation, New York, pp 301–322

Raudenbush SW (2009) Statistically analyzing effect sizes: Random effects models, 2nd edn. Russell Sage Foundation, New York

Robinson D (1995) The impact of cognitive skills training on postrelease recidivism among Canadian federal offenders. Correctional Research and Development, The Correctional Service of Canada, Ottawa, ON

Rosenthal R (1991) Meta-analytic procedures for social research, Rev. ed edn. Applied social research methods series. Sage Publications, Newbury Park

Ross RR, Fabiano EA, Ewles CD (1988) Reasoning and Rehabilitation. Int J Offender Ther Comp Criminol 32:29-36

Schmidt F, Le H, Oh IS (2009) Correcting for the distorting effects of study artifacts in meta-analysis, 2nd edn. Russell Sage Foundation, New York

Schulze R (2004) Meta-analysis: A comparison of approaches. Hogrefe & Huber, Toronto

Shadish WR (2007) A world without meta-analysis. J Exp Criminol 3:281–291. doi: 10.1007/s11292-007-9034-0, URL http://dx.doi.org/10.1007/s11292-007-9034-0

Shadish WR, Haddock CK (2009) Combining estimates of effect size, 2nd edn. Russell Sage Foundation, New York

Sánchez-Meca J, Marín-Martínez F, Chacón-Moscoso S (2003) Effect-size indices for dichotomized outcomes in meta-analysis. Psychol Methods 8:448–467, URL jsmeca@um.es

Stern JM, Simes RJ (1997) Publication bias: Evidence of delayed publication in a cohort study of clinical research projects. BMJ 315:640–645

Sterne JAC, Becker BJ, Egger M (2005) The funnel plot. Wiley, Chichester, pp 75–99

Viechtbauer W (2005) Bias and efficiency of meta-analytic variance estimators in the random-effects model. J Educ Behav Stat 30:261–293. doi: 10.3102/10769986030003261, URL http://jeb.sagepub.com/cgi/content/abstract/30/3/261

Weisburd D, Lum CM, Yang SM (2003) When can we conclude that treatments or programs "don't work"? Ann Am Acad Pol Soc Sci 587:31–48, doi: 10.1177/0002716202250782, URL http://ann.sagepub.com/cgi/content/abstract/587/1/31

Wilson DB (2009) Missing a critical piece of the pie: Simple document search strategies inadequate for systematic reviews. J Exp Criminol. doi: 10.1007/s11292-009-9085-5

Wilson DB, Bouffard LA, Mackenzie DL (2005) A quantitative review of structured, group-oriented, cognitive-behavioral programs for offenders. Crim Justice Behav 32:172–204. doi: 10.1177/0093854804272889, URL http://cjb.sagepub.com/cgi/content/abstract/32/2/172