
Alison J. O’Mara, Herbert W. Marsh, and Rhonda G. Craven

SELF Research Centre, University of Western Sydney, Australia

Meta-analysis is a valuable statistical technique for synthesising the available educational research literature on a particular topic, owing to its comprehensive and relatively unbiased approach to analysis. Since its conception by Glass in 1976, meta-analysis has been embraced by many researchers in education, psychology, and other disciplines, leading to the evolution of several very distinct methods for conducting meta-analysis – each with their own unique implications. The present paper starts with a discussion of the purpose and value of meta-analysis. This is followed by an overview of the features of the most common meta-analytic techniques (fixed effects models and random effects models), and the emerging technique of multilevel modeling meta-analysis. Historical and contemporary issues in meta-analysis are detailed, with particular attention to publication bias, generalisability and the assumptions inherent in various meta-analytic techniques, multivariate analyses and the independence of effect sizes, and power. Suggestions for how to address these issues will be provided. Finally, guidelines for selecting appropriate meta-analytic methods will be presented, as well as suggested resources for researchers wishing to conduct a meta-analysis. The emphasis of the paper is on understanding the various approaches to conducting responsible meta-analysis and their implications, rather than providing a prescriptive account of performing the technique.

What is a Meta-Analysis?

It is becoming increasingly evident that major scientific issues cannot be resolved by a single study, and that advances in knowledge come from the integration of many studies (Schmidt, 1992). Thus, the purpose of meta-analysis is to synthesise results from multiple studies to observe effect sizes across those studies on the phenomenon under review (Rosenthal, 1984), with the aim of gaining a greater understanding of related research reports. Meta-analyses approach research similarly to the more traditional experimental or correlational methods. The steps in a typical meta-analysis include data collection, evaluation, analysis, interpretation, significance testing, and drawing conclusions, just as in other research types. Hence, meta-analysis has been called a primary research investigation in itself with unique innovative characteristics in relation to research design (Cooper, 1998). Perhaps the most obvious difference between meta-analysis and other primary research techniques is that studies (rather than people) are treated as the unit of analysis.

Traditional methods of literature review focus on statistical significance testing, which is problematic because significance testing is highly dependent on sample size. Meta-analysis changes the focus from significance to the direction and magnitude of the effects across studies. This is achieved through the calculation and analysis of effect sizes, which standardise the findings across studies such that they can be directly compared. In this sense, the effect size is the dependent variable. It is the unique quantification of research findings in meta-analysis that makes it a powerful synthesis technique; but it is also in the calculation, analysis and interpretation of effect sizes that meta-analysis can be (accidentally or intentionally) misused. After a brief summary of the benefits of meta-analysis, we will outline a few common methods of conducting a meta-analysis, and the issues and considerations inherent in each method.

The Value of Meta-Analytic Techniques

Meta-analysis is an increasingly popular research technique, and for good reason. Meta-analysis is a replicable and defensible method of synthesising findings across studies. This is achieved through the guidelines imposed on how one can conduct a meta-analysis, thus instilling discipline in the research synthesis process. Meta-analyses go beyond traditional research reviews in that they can identify patterns that are regularly veiled by traditional null-hypothesis testing (Schmidt, 1992) because they organise information found in research reports in a meaningful way (Lipsey & Wilson, 2001), and have more statistical power than narrative reviews (Cohn & Becker, 2003). Further, meta-analyses are useful in identifying trends when sample sizes are small, which is common in areas of education, such as the gifted and talented literature (Asher, 2003).
Meta-analysis can also handle the synthesis of a large number of studies, which may not be feasible in a conventional literature review. As a result, meta-analyses are particularly useful when research in a field produces confounding or vague results (Wolf, 1986), or when there are large numbers of studies on the topic of interest. By facilitating generalisation of the knowledge gained through individual studies, meta-analysis promotes replication of research. Meta-analysis can also help to identify gaps in the research literature by indicating which information was insufficiently reported in the literature. In this way, it can establish a solid foundation for the next generation of research on that topic.

Importantly, the systematic review of the literature in meta-analysis allows more refined assessment of the author’s hypotheses, methods, results and conclusions (Lipsey & Wilson, 2001). That is, the findings are represented in a more sophisticated way than in traditional literature reviews, which can help to minimise bias and making spurious conclusions based on imprecise overviews. As such, meta-analysis is frequently considered to be more objective than traditional literature reviews (Cooper & Hedges, 1994b). Meta-analysis can also account for the context of the research, including cultural, historical and social factors that may have impacted upon the research results. This helps to protect against making over-evaluations of the differences between studies. In these and other ways, meta-analysis is a valuable enterprise.

Why Meta-Analyses of Educational Research are Warranted

The literature on educational phenomena is diverse in both its foci and its methods. The foci of educational researchers include academic abilities, physical activity, creative expression, personal development, and cultural awareness, among others. The methods include intervention programs, behavioural observations, standardised testing, and survey data collection. This diversity is largely due to the importance of educational settings in the development of human thought, socialisation and enterprise, which motivates careful, multiple-approach investigation. Meta-analysis can help to organise this diverse literature, thereby making sense of the findings. Additionally, meta-analysis helps to overcome the problem of small sample sizes that plagues many areas of educational psychology, such as gifted and talented programming (Asher, 2003).

From the perspective of output, meta-analysis has the ability to produce concise, easily interpretable findings regarding conceptual issues, program effectiveness, and research design strategies. That is, a single meta-analysis has the capacity to inform and appeal to educational theoreticians, practitioners, policymakers, and researchers alike. This in turn can promote a consonance within the (rather segregated) educational community, potentially leading to more effective, concentrated efforts. For these reasons, and also for the overall value of meta-analysis discussed above, meta-analysis can make a significant contribution to educational research.

The Widespread Contribution of Meta-analysis to Education

To illustrate both the broad subject-area applicability and the potential impact upon future research of meta-analyses in education, below is a summary of the five most highly cited meta-analyses that have been published in Review of Educational Research (according to ISI Web of Science). The number of citations indicates how many times that meta-analysis has been cited in papers in the ISI database: classroom structure (Cohen, 1994; 184 citations), reinforcement and motivation (Cameron & Pierce, 1994; 138 citations), Literacy (Bus, Vanijzendoorn, & Pellegrini, 1995; 115 citations), Sex-differences and intellectual abilities (Feingold, 1992; 78 citations), school learning (Wang, Haertel, & Walberg, 1993; 75 citations). The diversity of the topics covered and the frequency of citation rates demonstrate the relevance and usefulness of meta-analysis for education researchers.

Meta-analytic Methods

There are many phases in a meta-analytic review. The starting point is to define the area of interest, specify criteria for inclusion in the meta-analysis, and then collect the studies. One must also develop coding materials that will be used to collect the information on variables of interest, such as participant characteristics. The early stages of these procedures are detailed elsewhere (see for example Lipsey & Wilson, 2001; Cooper & Hedges, 1994a). We start our discussion of meta-analytic procedures at the calculation of effect size stage, and proceed through to data analysis.
**Computation of Effect Sizes**

The effect size essentially encodes the outcomes of the selected research findings into a numeric value and is the dependent variable in the analyses. Effect sizes have several distinctive characteristics that make them pertinent to the meta-analytic process: they must be comparable across studies (which generally require standardization); represent the magnitude and direction of the relationship of interest; be independent of sample size; and have a calculable standard error. There are two main types of effect sizes: those based on mean differences (commonly called \( d \)) and those based on correlations (commonly referred to as \( r \)). Each effect size type may have multiple methods of computation. Only the most commonly used will be discussed here (for further discussion see Lipsey & Wilson, 2001; Cooper & Hedges, 1994a).

### Mean Differences

Effect sizes based on mean differences (\( d \)) look to compare the average difference on a particular variable between two groups or between two time frames. Thus, there are two sub-types of mean differences (Lipsey & Wilson, 2001). The first is group contrasts, such as comparing male and female scores, or treatment group with control group scores on a particular variable. In education, psychology, and the social sciences, this is most commonly calculated using the standardized mean difference. The second type of mean difference effect sizes are based on pretest-posttest contrasts, which are used to measure changes in a variable over time. This is most typically calculated using the standardized mean gain (note that it is also possible to conduct a meta-analysis of raw mean differences rather than standardised mean differences, although this is much less common; see Bond, Wittala, & Richard, 2003).

The standardised mean difference (\( d \)) can be calculated by inputting means and standard deviations, \( t \)-values, \( F \)-values, or \( p \)-values (see Lipsey & Wilson, 2001; Wolf, 1986). The most common calculation is through means and standard deviations, which is represented by the formula \( d = (M_t - M_c) / SD_{pooled} \), where \( M_t \) is the mean of the treatment group (or posttest score in the standardised mean gain), \( M_c \) is the mean of the control group (or pretest score in the standardised mean gain), and \( SD_{pooled} \) is the pooled standard deviation of both groups. In this way, a positive effect size suggests that the treatment group had a larger improvement than the control group, whereas a negative \( d \) value indicates a higher score for the control group compared to the treatment group.

Mean difference effect sizes typically range in magnitude form 0 to 1, although it is not uncommon to have effect sizes that are as large as 5 in magnitude. Cohen’s (1988) rule of thumb suggests the following guidelines for interpreting the meaningfulness of the magnitude of the observed standardized mean difference effect size: if \( d \leq 0.20 \) the effect is small; if \( d = 0.50 \) the effect is medium or moderate; if \( d \geq 0.80 \) the effect is large.

### Correlations

Effect sizes based on correlations compare the relative strength of the association between two variables. For instance, one could have an effect size that reflects the correlation between age and academic self-concept, or between socioeconomic status and achievement test scores. The reliability of a measure, as indicated by a correlation between test scores, is also readily recognised type of correlation that may be used in meta-analysis. The product-moment correlation is used to determine the relation between two continuous variables, the point-biserial correlation coefficient is used for the relation between a continuous and a dichotomous variable, while the relation between two dichotomous variables is calculated through the odds-ratio or the phi coefficient (Lipsey & Wilson, 2001).

The effect size for the relation between two continuous variables (the product-moment correlation, \( r \)) is calculated by dividing the covariation between variables \( x \) and \( y \) (\( \sigma_{xy} \)) by the product of the standard deviations of the variables \( x \) and \( y \) (\( \sigma_x \) and \( \sigma_y \), respectively). The formula is therefore represented by \( r = \sigma_{xy} / (\sigma_x \sigma_y) \). A large \( r \) represents a strong correlation between the two variables. For computational formulae of the other correlation effect sizes, see Lipsey and Wilson (2001). Correlation coefficient effect sizes range in magnitude between 0 and 1, with 1 being an exact relation between the variables, and 0 denoting no relation at all. Cohen’s (1988) rule of thumb for the correlation coefficient is as follows: if \( r \leq 0.10 \) the effect is small; if \( r = 0.25 \) the effect is medium or moderate; if \( r \geq 0.40 \) the effect is large.
Fixed Effects Models

Traditionally, meta-analyses have utilized fixed effects models. A core assumption of the fixed effects model is that all of the variability between effect sizes is due to sampling error alone. There are two main strategies in fixed effect models. The first is homogeneity testing, also known as categorical fixed effects model testing, or the meta-analytic analogy to the ANOVA. This strategy looks for systematic differences between groups or categories of responses (e.g., primary schools, junior high schools, and senior high schools) within a variable (e.g., school level). This technique is appropriate for categorical variables. The second strategy is a weighted multiple regression, typically utilising backward elimination. This strategy tests the ability of categories within each variable to predict the effect size. It is appropriate when the data to be analysed are continuous variables and/or there are multiple variables to be analysed.

Homogeneity Testing

Homogeneity analyses test the assumption that all of the effect sizes are estimating the same population. For categorical variables, meta-analysts use categorical models to ascertain the association between the study features and the extent of effect sizes (Hattie, Biggs & Purdie, 1996). “These models provide a between-classes effect (analogous to a main effect in an ANOVA design) and a test of homogeneity of the effect sizes within each class” (Hattie et al., 1996, p. 112).

Firstly, the fixed effects model weights each study by the inverse of the sampling variance. Then the homogeneity within-group and between-group statistics are calculated using an analog to the ANOVA (see Lipsey & Wilson, 2001). The within-group homogeneity statistic determines whether each set of outcomes has an effect size consistent across the studies. That is, the within-group statistic reflects whether the studies within each cell of the particular variable are homogeneous. This allows determination of the appropriateness of the grouping of studies for between-group analysis. Ideally, within-group findings should be above the .05 probability level (Lipsey & Wilson, 2001). The between-group homogeneity statistic is used to estimate the between-groups effect, which reveals whether the average effect size differs between groups. That is, the between-group statistic indicates whether the variable is a significant moderator of outcome, and should be below the .05 significance level. Thus, a significant between-group and a non-significant within-group statistic are indicative of a well-defined model. If homogeneity is rejected, there are real between-study differences and the studies therefore must estimate different population mean effect sizes. There are programs that can conduct the analog to the ANOVA for meta-analysis (e.g., Comprehensive Meta-analysis, Borenstein & Rothstein, 1999); alternatively, David Wilson has developed easy to use macros that can be implemented with SPSS, Stata and SAS (Lipsey & Wilson, 2001).

Multiple Regression

Regression is appropriate when there are continuous variables (e.g., age) and/or if there are multiple variables that are valuable to consider simultaneously (e.g., both age and gender). That is, unlike in the homogeneity analyses, the regression can suggest the optimal combination of moderators to predict the effect size outcomes. Weighted least squares multiple regression are recommended in these instances (Lipsey & Wilson, 2001), and regression in meta-analysis is always weighted to ensure that differences in sample size are accounted for (Durlak & Lipsey, 1991). A significant Q residual from the fixed effects weighted multiple regression analysis suggests homogeneity in the category; that is, there is within-group variance (Lipsey & Wilson, 2001). Although common statistical packages (e.g., SPSS) can conduct multiple regression, they can report incorrect standard errors; as such, David Wilson has also developed macros for conducting a multiple regression that can be implemented with SPSS, Stata and SAS (Lipsey & Wilson, 2001).

Random Effects Models

Random effects models are becoming more prevalent in meta-analytic research. A central assumption of the random effects model is that the variability between effect sizes is due to sampling error (as in fixed effects models) plus variability in the population of effects. That is, features of the studies themselves contribute to the variability in effect sizes in addition to the random sampling error. Such models follow the same procedures as fixed effects models (i.e., homogeneity analyses and regression), except that they add a random variance
component to the variance when conducting the analyses. By accounting for this additional variance component, the random effects model allows generalisability to the greater population of studies that may not have been included in the meta-analysis for whatever reason. If the within-group homogeneity statistic in the fixed effects analog to the ANOVA or the Q residual from the fixed effects weighted multiple regression analysis is significant, then the random effects model should most definitely be used (Lipsey & Wilson, 2001).

Thus, the random effects model weights each study by the inverse of the sampling variance plus a constant variance component representing the variability across the population effects. In other words, the random effects variance component is added to the variance associated with each effect size, and this total is inverted to produce the weighting. The variance component is typically calculated as

\[ \theta = \frac{Q - (k - 1)}{\sum w_i - \left( \frac{\sum w_i^2}{\sum w_i} \right) \cdot \left( \frac{\sum w_i^2}{\sum w_i} - 1 \right)} \]

where \( Q \) is the homogeneity statistic produced in the analog to the ANOVA, \( k \) is the number of effect sizes, and \( w_i \) is the weight of effect size \( i \). The analyses are then conducted with this new weighting.

Random effects models are more conservative than fixed effects models. This is manifested where effects deemed to be significant under a fixed effects model may not be significant under a random effects model, and random effects models typically exhibit larger confidence intervals. When homogeneity is maintained, the two methods (fixed and random effects models) produce similar results (Hedges & Vevea, 1998).

**Multilevel Modeling and Meta-Analysis**

*The Hierarchical Nature of Meta-Analysis and the Benefits of Multilevel Meta-Analyses*

Meta-analytic data is inherently hierarchical. One will find any or all of the following conditions in their research: multiple outcomes per study, multiple measurement points per study, and multiple sub-samples per study. These factors typically result in multiple effect sizes per study, which are not independent (as discussed above). Any specific analysis can only include one effect size per study (or one effect size per sub-sample within a study), and at such, analyses almost always are of a subset of coded effect sizes. Given that effect sizes are embedded in this hierarchical structure (e.g., effect sizes nested within participants, which are nested within studies), multilevel modelling can be applied to meta-analysis (Goldstein, 1995; Hox, 2002; Bryk & Raudenbush, 1992). Indeed, meta-analysis is increasingly being seen as a special form of multilevel regression analysis (Hox & de Leeuw, 2003).

There are multiple benefits of conducting a meta-analysis using multilevel modelling. Multilevel modeling can account for both sample variance and systematic variance in mean effect size estimation (Severiens & ten Dam, 1998), and in this way it can account for large variation between studies (Swanborn & de Glopper, 1999). This also means that multilevel meta-analysis allows generalisation of the findings to the population since it incorporates random effects. Multilevel modelling also allows the extension of traditional random effects models to include covariates (Van den Noortgate & Onghena, 2003; Goldstein, Yang, Omar, Turner, & Thompson, 2000), provides more precise and less biased estimates of between-study variance than traditional techniques (based on Hedges and Olkin’s formula) (Van den Noortgate & Onghena, 2003), and increases the accuracy of the estimation of standard errors on parameter estimates and the assessment of the significance of explanatory variables through improved modelling of the nesting of levels within studies (Bateman & Jones, 2003). Circumvent the issue of independence that plagues meta-analytic research (Kalaian & Raudenbush, 1996). Multilevel modelling accounts for dependencies in the data whilst also allowing each study to contribute different effect sizes, thus maximizing the amount of information included in the analyses (Dickerson & Kemeny, 2004). As such, it is possible with multilevel modelling to specify a multivariate outcome model (Hox & de Leeuw, 2003). In summary, there are numerous advantages of using multilevel meta-analyses over traditional meta-analytic techniques.

**So Where are all the Multilevel Model Meta-Analyses?**

Despite the growing appreciation of the ‘natural’ relation between multilevel models and meta-analysis (Draper, 1995), traditional meta-analytic practices are still commonly used (Van den Noortgate & Onghena, 2003). Although the notion of approaching meta-analysis from a multilevel modelling perspective has been around since Raudenbush and Bryk (1985) first posited it two decades ago, a literature search on meta-analyses that have been conducted using multilevel modelling reveals only a handful of studies. The lack of published
meta-analyses utilizing this method does not stem from scepticism about its potential efficacy; rather, a distinct lack of prescriptive instructions on how to conduct a meta-analysis using multilevel modelling has made researchers hesitant (Van den Noortgate & Onghena, 2003). Only recently have more detailed methodological notes been published (e.g., Hox & de Leeuw, 2003; Van den Noortgate & Onghena, 2003).

Conducting a Multilevel Model Meta-Analysis

As mentioned above, there is limited information on how to conduct a multilevel modeling meta-analysis available. The website for the ATS Stat Consulting Group at UCLA (http://statcomp.ats.ucla.edu/mlm/) provides excellent examples on how to conduct multilevel modeling meta-analysis in HLM, MLwiN, and Stata to be used in conjunction with Hox’s (2002) book. However, they do not include examples for multivariate outcomes.

The first step in the multilevel model meta-analysis is to estimate the intercept-only model. For this, no predictors are included in the model. Multilevel models in meta-analysis stipulate two equations: An outcome level component (level 1), which predicts the effect size across the self-concept outcome domains, and a study-level component (level 2), which predicts the effect sizes across the studies. The outcome-level component is represented by the level 1 equation \( d_j = \delta_j + e_j \) and the study-level component is represented by the level 2 equation \( \delta_j = \gamma_0 + u_j \). In the level 1 equation, \( d_j \) refers to the effect size from study \( j \), \( \delta_j \) refers to the intercept (average effect size for an average outcome), and \( e_j \) refers to the random error (residual) at level 1. For the level 2 equation, \( \gamma_0 \) represents the regression coefficient, and \( u_j \) is the level 2 random error. The level 1 model is the same as the random effects model (Hox, 2002), as it incorporates both the outcome-level and the study-level components.

The variance of \( u_j \) (the residual error term) indicates the variability in effect sizes across studies. If the studies are homogeneous (i.e., there is no significant between-study variance), there is no significant difference between studies on the variable of interest. Any variance in effect sizes is due purely to sampling variance (Lipsey & Wilson, 2001). As such, no further analyses need to be conducted. However, a significant variance component suggests that there is variance unexplained by the model: The model is heterogeneous (Bryk & Raudenbush, 1992). In other words, the participants take on a different form for a particular variable in each sample.

The following hierarchical model then applies:

\[ d_j = \gamma_0 + \gamma_1 W_{1j} + \gamma_2 W_{2j} + \ldots + \gamma_s W_{sj} + u_j + e_j \]

where \( d \) is the mean effect size, \( \gamma_0 \ldots \gamma_s \) are the regression coefficients, \( W_{1j} \ldots W_{sj} \) are the study characteristics (predictor or moderator variables, such as use of random assignment to treatment conditions), \( u_j \) is the systematic variability in study \( j \) not captured by the \( s \) predictors, and \( e_j \) is the sampling error for study \( j \) (Bryk & Raudenbush, 1992). The intercept (\( \gamma_0 \)) is the estimated effect size for a study with zero values for all moderator variables. The remaining regression weights (\( \gamma_0 \ldots \gamma_s \)) indicate the amount of expected variation in the effect size for a one-unit change on each variable.

Issues in Meta-Analysis and How to Address Them

Meta-analyses are a concentrated effort to synthesise and evaluate the available literature on a given topic. Although we encourage educational researchers to undertake meta-analyses, the decision to undertake one should not be taken lightly. There are a few caveats that need to be heeded before any meta-analysis is attempted, as this section attempts to highlight.

General Things to Consider Before Embarking Upon a Meta-analysis

There are a few general issues to consider before undertaking a meta-analysis. Firstly, meta-analyses are time-consuming and labour-intensive: the larger the field of interest, the longer it will take. Make sure you have plenty of time to collect the relevant studies (including any ‘fugitive studies’ if necessary), to develop the coding materials, to code the studies, and to run the extensive analyses. Secondly, the coding procedure often requires somewhat subjective interpretations of the study. You should have at least one other coder who is familiar with the area under investigation for inter-rater reliability purposes. It is a good idea to identify a research assistant
with appropriate expertise before you start the meta-analysis. It is also useful to involve the research assistant in the development of coding materials to ensure that they fully understand how the materials work and what they mean. On this point, if they are not involved in the code material development, you will need to train them in the interpretation of the codes. A few rounds of pilot coding are recommended to guarantee understanding and agreement amongst coders and enable the development of a detailed coding book.

Thirdly, it is important to remember that the selection criteria used can greatly affect your results, so stipulate them carefully. If you choose to only use published studies, for example, think about how this affects the generalisability of your results. Or if you restrict the studies to certain age groups, consider how the restriction limits the applicability of your findings. Fourthly, qualitative distinctions between studies may not be fully captured in the meta-analytic process. Consider whether this will negatively impact the usefulness of your findings for your particular topic.

Finally, The analysis of between study differences is fundamentally correlational, which has implications for causality. Do not make the mistake of assuming that, because of increased power and sample size, meta-analytic results actually prove anything – meta-analysis can only identify trends that may have otherwise been overlooked. This in itself is a very useful outcome, but meta-analysis is not the perfect solution to every problem.

### The Quality of Studies Synthesised

Eysenck (1994) published a rather critical treatise on meta-analysis. The primary concern that he noted was that the meta-analytic process did not allow any subjective evaluation of the quality of the studies or the instruments used within them. He argued that meta-analysis is flawed because “the computer [the tool of the meta-analyst] avoids the bias of the subjective approach but simply adds together the biases of the authors of the original reports – which may or may not balance out” (p. 789). Eysenck, as have others before and after him, have pondered the ramifications of including poor quality studies in a meta-analysis, because they have the potential to distort the results of the meta-analysis (Schulz, Chalmers, Hayes, & Altman, 1995).

The main way in which the quality of a study may be compromised is through the way in which the studies are designed (Hunter & Schmidt, 1990), such as the use of non-randomised group assignment and high participant attrition. A recommended tactic is to code the quality of the study and include it in the analyses to see if it does impact effect size (Wolf, 1986). Wortman (1994) details methods for assessing quality and detecting sources of bias, which may aid in this endeavour. It is interesting to note that, despite the widespread concern about including low quality studies in the meta-analysis, that meta-analyses often report no correlation between study quality and effect size (e.g., Lipsey & Wilson, 1993).

Another perhaps more easily identified way in which low quality studies can influence a meta-analyst is through the poor reporting of variables. Many studies fail to report information that may be of value to researchers, such as participant demographics, attrition rates, and even the specific results (means, standard deviations, etc.) when findings are non-significant. This can restrict the types of analyses the meta-analyst can conduct.

### Publication Bias

Publication bias is defined as “a bias against negative findings on the part of those involved in deciding whether to publish a study” (Soeken & Sripusanapan, 2003, p. 57). Given the apparent prevalence of positive findings in the published literature, meta-analyses are more likely to identify and include studies with positive outcomes, thereby potentially skewing meta-analytic findings towards a positive mean effect size. In an analysis of 48 published meta-analyses, Sutton, Duval, Tweedie, Abrams, and Jones (2005) found that (54%) of reviews had missing studies and that “around 5-10% of meta-analyses may be interpreted incorrectly because of publication bias” (p. 1576). The problem of publication bias is consequential.

Since the identification of the potential for publication bias, meta-analysts have developed various ways of accounting for publication bias (see comparisons of various techniques in Preston, Ashby, & Smyth, 2004; Soeken & Sripusanapan, 2003). Perhaps the most defensible technique is the trim and fill method (Duval & Tweedie, 2000a; Duval & Tweedie, 2000b). The primary benefit of the trim and fill algorithm is that it both evaluates and adjusts for the possibility of publication bias, whereas other techniques (e.g., Rosenthal’s Fail Safe
N) only estimate the number of missing studies or the extent of the publication bias problem without addressing the issue. This method is being increasingly adopted in reputable journals.

**Aggregation Across Disparate Studies**

An oft-cited argument against meta-analysis is that it aggregates data from disparate types of studies (Eysenck, 1994; Wolf, 1986); that meta-analysis tries to compare ‘apples and oranges’. This has been addressed by the development of stronger meta-analytic techniques such as random effects homogeneity testing (Lipsey & Wilson, 2001) and multilevel modelling methods (Swanborn & de Glopper, 1999), in which it can be determined whether categories within a variable actually come from the same population, the results of which are generalisable to the studies not included in the meta-analysis. Meta-analysis, through the inclusion of variance component estimates, is able to account for differences between studies in a way that traditional literature reviews fail to do. However, Cortina (2003) suggested that the search for moderators through homogeneity testing is inconsistent amongst meta-analysts, which can compromise the synthesis of non-identical studies and thus the generalisability of the results. Cortina therefore developed guidelines for conducting moderator analyses, making the ‘apples and oranges’ debate less contentious. As with all research, it is the responsibility of the meta-analysts to judiciously evaluate and report the results of the meta-analysis; a task which is arguably less prone to bias than a traditional literature review.

**Homogeneity Testing: Assumptions and Generalisability**

An essential aspect of any meta-analysis is to test for the homogeneity of the effect sizes. A homogeneous distribution of effect sizes means that the dispersion of the effect sizes around their mean is no greater than that brought about by sampling error. In other words, the results across studies are consistent, and the mean effect size and its standard error summarises the relation between the independent variable and their outcome/s (Kalaja & Raudenbush, 1996). A heterogeneous distribution means that the variability of the effect sizes is larger than that expected by sampling error and the differences between effect sizes have some other cause, which leads us to search for potential explanatory variables. Once we have established heterogeneity, it is important to run moderator analyses to identify the variables that may explain the differences between effect sizes. This is generally achieved through an analog to the ANOVA known as a homogeneity or Q-test.

Traditionally, homogeneity tests were conducted from a fixed effects model approach. An implicit assumption of the fixed effects model is that any random error in the effect size distribution is derived only from subject-level sampling error (Hedges & Vevea, 1998). That is, fixed effects models assume that all the included studies drew their samples from the same population, and that there is only one true population effect (Cohn & Becker, 2003; Erez, Bloom, & Wells, 1996). This is increasingly being recognised as an unrealistic supposition, as features of the studies themselves (such as the treatment setting) may also be a source of variance (Raudenbush, 1994). Further, if the assumption of homogeneity is not met, fixed effects models may underestimate the error variance (Overton, 1998).

The random effects model attempts to rectify this by investigating how the between-study differences affect the relation between the independent variable/s and the study outcomes. This model assumes that studies are heterogeneous to an extent (Erez et al., 1996), because each study has different contexts, researchers, and even methods. By accounting for variance beyond the subject-level error, the random effects model is more generalisable to the larger population of studies (Raudenbush, 1994). In other words, random effects models allow inferences to be drawn to studies beyond the sample used (Valentine et al., 2004; Cohn & Becker, 2003; Hedges & Pigott, 2001; Hedges & Vevea, 1998). However, Overton (1998) suggests that random effects models may overestimate the error variance if homogeneity is not met, which means that in some instances the random effects model may be too conservative.

The relatively new approach of multilevel meta-analysis incorporates meta-analysis into a multilevel modeling structure. It also allows generalisation of the findings to the greater population of studies as it incorporates a variance component in much the same way that random effects models do (Van den Noortgate & Onghena, 2003). If one is purely seeking to evaluate the characteristics of the sample, or if one has reason to believe that their sample constitutes the entire population, then the fixed effects model is appropriate; in all other
cases were generalisability is desired, then a random effects model or a multilevel model is more appropriate (Lipsey & Wilson, 2001).

**Multivariate Analyses and the Independence of Effect Sizes**

Violations of independence occur when studies produce multiple effect sizes because of multiple treatment groups or multiple outcome measures. Effect sizes from the same study are likely to be correlated in a way that may distort the results of the meta-analysis (Cooper, 1998). Thus, meta-analyst must establish an independent set of effect sizes before proceeding with the analysis.

This can be achieved by conducting separate analyses on each outcome. However, this may not always make sense for the research question at hand. Consider a study by O’Mara, Marsh, Craven and Debus (in press), which sought to establish the importance of a construct validity approach to self-concept intervention research. The researchers wanted to distinguish whether a particular self-concept domain outcome that was focally relevant to the intervention would yield a higher effect size than a self-concept domain that was less relevant. For instance, an intervention designed to increase participants’ physical ability self-concept should exhibit higher effect sizes on a measure of physical ability self-concept than on, say, social self-concept. If a study reported both physical ability self-concept and social self-concept scores, these scores are not independent because they have many features in common such as the same intervention, the same administrator, etc, and are therefore more correlated than effect sizes from other studies. And yet the central question of the study required that they compare multiple effect sizes from the same study. In cases such as this, a compromise is to use Cooper’s (1998) shifting unit of analysis technique. In the example at hand, the effect size for the two self-concept domains would initially be averaged to produce a single effect size for calculations involving the overall effect size for the sample. That is, the effect sizes are aggregated to produce one effect size for each intervention. For each moderator analysis, effect sizes are aggregated based upon the particular moderator variable, such that each study only includes one effect size per outcome on that particular variable. Although this does not eliminate the problem of independence, this approach minimizes violations of assumptions about the independence of effect sizes, whilst preserving as much of the data as possible (Cooper, 1998). An exciting new direction for dealing with multiple effect sizes is to use a multilevel model approach, which will be discussed at length below.

**Statistical Power**

Statistical power is the probability that a test accurately rejects a false null hypothesis, also known as Type II error (Rosnow & Rosenthal, 1996; Becker, 1994). Power is increasingly seen as an important consideration in research using statistical analyses, as the rather arbitrary .05 criterion can lead to mistaken conclusions (Rosnow & Rosenthal, 1996). Generally, meta-analyses have high statistical power (Miller & Pollock, 1994), which is considered to be one of the major advantages of the technique (although this has been contended, see Hedges & Pigott, 2001). Cohn and Becker (2003) emphasised the importance of this, stating that

Traditional narrative reviews are unable to distinguish between null findings that result from low power and null findings that reflect a genuine absence of population effects, sometimes leading reviewers to erroneously conclude that a body of evidence does not support a proposed relationship or treatment effect (p. 244).

There are a few issues related to the choice of meta-analytic model when considering power. In fixed effects meta-analyses, increasing the sample size can increase power; this does not necessarily hold true for random effects models (Cohn & Becker, 2003). Indeed, in random effects models, increasing the number of studies may lead to a larger variance component compared to using a smaller sample of studies, which can increase the standard error and thereby reduce statistical power. However, relative to the fixed effects model, random effects meta-analyses generally reduce the standard error of the weighted mean effect size, and so are generally seen as providing greater power (Cohn & Becker, 2003). Hedges and Pigott (2001) also suggest that excluding studies with very small sample sizes from a random effects meta-analysis may increase power because of the heterogeneity such studies usually bring to the meta-analysis; although any decision to systematically exclude studies should be well-considered.
Guidelines for Selecting Appropriate Meta-Analytic Methods

A number of things need to be taken into account when selecting a meta-analytic model. Obviously, the inferences one wishes to make (i.e., is it important to be able to generalise to studies not in the population) is very important. The type of independent variable/s and the number of dependent variables are also key factors. Issues such as power, the threat of publication bias, accessibility to appropriate statistical programs, and the statistical confidence of the meta-analyst may also influence the decision. Table 1 represents a very simple decision guide for meta-analysts, although clearly a responsible meta-analyst must factor in all considerations discussed in this paper and elsewhere (e.g., Lipsey & Wilson, 2001; Cooper & Hedges, 1994a) before selecting a model. Note that in some instances, more than one model type may be appropriate for a particular situation.

Table 1. Guidelines for Selecting Appropriate Meta-Analytic Models

<table>
<thead>
<tr>
<th>Number of Dependent Variable/s</th>
<th>Type of independent variable/s</th>
<th>Generalisability required?</th>
<th>Suggested appropriate method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Categorical</td>
<td>No</td>
<td>Fixed effects analog to the ANOVA</td>
</tr>
<tr>
<td>1</td>
<td>Continuous or multiple</td>
<td>No</td>
<td>Fixed effects weighted multiple regression</td>
</tr>
<tr>
<td></td>
<td>dichotomised variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Categorical</td>
<td>Yes</td>
<td>Random effects analog to the ANOVA</td>
</tr>
<tr>
<td>1</td>
<td>Continuous or multiple</td>
<td>Yes</td>
<td>Random effects weighted multiple regression or multilevel modeling</td>
</tr>
<tr>
<td></td>
<td>dichotomised variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;1 (multivariate)</td>
<td>Categorical</td>
<td>Yes</td>
<td>Random effects analog to the ANOVA using Cooper’s Shifting Unit of Analysis Technique</td>
</tr>
<tr>
<td>&gt;1 (multivariate)</td>
<td>Continuous or multiple</td>
<td>Yes</td>
<td>Multilevel modeling</td>
</tr>
<tr>
<td></td>
<td>dichotomised variables</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Suggested Starter Resources

An excellent starting point for conducting any meta-analysis is Lipsey and Wilson’s (2001) book, *Practical meta-analysis*. The book runs through the steps in a meta-analysis, analytical issues, interpretation issues, and easy to follow examples. Cooper and Hedges (1994a) present a collection of papers from leading meta-analysts detailing everything from data collection to model specification to results presentation. This is a more detailed companion to the Lipsey and Wilson book. Hedges and Olkin’s (1985) seminal work provides insight into the development of fixed and random effects models, and explicates the statistical logic behind these models.

For those who are considering multilevel modeling, good introductory chapters can be found in Hox (2002) and Hox and de Leeuw (2003). Van den Noortgate and Onghena (2003) provide a simple to understand comparison between traditional meta-analysis (i.e., fixed and random effects models) and multilevel meta-analysis. For a deeper, more technical understanding of multilevel meta-analysis, Raudenbush has published some informative works. Raudenbush was one of the pioneers who first called for multilevel model meta-analyses (see Bryk & Raudenbush, 1992 for a brief overview), and is one of the few to have written about multivariate multilevel meta-analysis (see Kalaian & Raudenbush, 1996; see also the brief section in chapter 3 of Goldstein, 1995).

Conclusion

Meta-analysis is an invaluable research tool. It has many benefits, but there are also several important conceptual and statistical considerations that any responsible researcher needs to evaluate before commencing a meta-analysis. We have attempted to elucidate the most contentious and contemporary issues in meta-analysis, with the dual purpose of informing meta-analytic practitioners as well as aiding their audience to digest the complexities of this developing field. It is hoped that this introduction to the issues will help newcomers to the
field feel better equipped to tackle decisions about the choice of models, publication bias, the independence of effect sizes, and power, and that readers of meta-analyses feel more confident in comprehending meta-analytic findings.

About the Authors

Alison O’Mara is a Doctoral candidate with the SELF Research Centre. Her research involves a series of multilevel model meta-analyses related to multidimensional self-concept. Alison has presented refereed papers at international conferences, is the past recipient of the National Leadership Scholarship, and the current recipient of a prestigious Australian Postgraduate Award.

Professor Herb Marsh is Director and Research Professor of the SELF Research Centre, University of Western Sydney. Largely due to his established research program into self-concept, motivation, identity and related constructs that are the focus of the SELF Research Centre, he is arguably Australia’s leading researcher in both the broad disciplines of Education and Psychology

Professor Rhonda Craven is Deputy Director of the SELF Research Centre – ranked 7th in the world in educational psychology, is Associate-Professor in the School of Education and Early Childhood Studies, University of Western Sydney. As an Educational Psychologist her research focuses on large-scale quantitative research studies in educational settings.

Contact Details

Ms. Alison O’Mara
PhD Candidate
SELF Research Centre
University of Western Sydney
Bankstown Campus
Locked Bag 1797
Penrith South DC NSW 1797
Australia
Email: a.omara@uws.edu.au
Phone: +61-414-726-535

References


