Ongoing developments in meta-analytic and quantitative synthesis methods: Broadening the types of research questions that can be addressed

Alison O'Mara-Eves and James Thomas

Social Science Research Unit, UCL Institute of Education, University College London, University of London, London, UK

Abstract

The purpose of this paper is to outline ongoing developments in meta-analytic methods and quantitative approaches to synthesising evidence. We discuss the increased awareness by meta-analysts of the need for methods which better grapple with complex social contexts, and meta-analysts' responses to the increasing informational needs of review audiences by developing methods are that are more fit for purpose and fit for use than their predecessors. Models of meta-analysis that we cover include both aggregative (e.g., classical meta-analysis) and configurative approaches (including subgroup analyses, meta-regression and multilevel analyses, multiple outcome analyses, and network meta-analysis). We then consider the role of additional data sources and multi-method approaches to synthesis by focusing on mixed methods synthesis, the use of largescale datasets and individual participant data, and qualitative comparative analysis. We highlight key issues for meta-analysis in educational research (publication bias and interpreting meta-analytic results). We end with reflections on the relation between meta-analytic methods and theory, and a discussion of how meta-analysis in education can move forward. Throughout, we place a particular emphasis on the importance of using a method that is appropriate for the research question, and how emerging methods allow us to address a broader range of research questions.

Key words
Meta-analysis, quantitative synthesis, research synthesis methods, systematic review, research methodology

Introduction
A quantitative method for synthesising quantitative evidence, meta-analysis is one of the most common and well-known of all methods for synthesising the findings from multiple research studies, usually within the context of a systematic review. Its use in educational research has been steadily growing since the term was coined by Glass in 1976; Figure 1 shows the number of educational research publications with 'meta-analysis' in the title in the Web of Knowledge database over the last 40 years[1]. Glass (2015) lightly summarised the history of meta-analysis as follows: "In 40 years, meta-analysis has grown from an obscure preoccupation of a small group of statisticians to a major academic industry" (p. 1).

Since its inception, meta-analytic methods have continued to be developed at an astounding rate, with new quantitative synthesis methods evolving, some of which resemble their meta-analysis cousin, whereas others are a new species altogether. At least three factors are driving this change: the realisation that classical meta-analytic methods do not allow us to address the increasingly complex questions being asked; that meta-analysis consumers often need more from evidence reviews than a single overall effect size; and that there is a role for meta-analyses in terms of theory generation as well as theory testing and describing the evidence base. Given the developments, it seems timely to review the state of the art of meta-analytic methods and consider emerging quantitative synthesis methods.

[1] Search conducted in Web of Knowledge on 3 August 2015 using the following strategy: TITLE: (meta-analy*) OR TITLE: (meta analy*) Refined by: RESEARCH AREAS: (EDUCATION EDUCATIONAL RESEARCH) Timespan: 1975-2014.

The purpose of this paper is to outline ongoing developments in meta-analytic methods and related methods for quantitative synthesis. A core theme throughout the paper is that meta-analysts and research synthesists are responding to the needs of their intended audiences—policymakers, decision-makers, and practitioners—by developing methods that enable more complex questions, complex phenomena, and complex contexts to be addressed. We start with a discussion of the distinction between aggregative and configurative approaches to synthesis. We then describe classical meta-analysis as an aggregative approach. Configurative approaches, in which we attempt to organise, explore, or explain trends and variation in the data, are described next: subgroup analyses, meta-regression and multilevel analyses, multiple outcome analyses, and indirect comparisons (or network meta-analysis). We then consider three multi-method approaches to synthesis: mixed methods synthesis, the use of large-scale datasets and individual participant data, and qualitative comparative analysis. Having covered the common and emerging methods for quantitative synthesis, we then turn to publication bias and interpreting meta-analytic results as key issues for meta-analysis in educational research. The paper finishes with reflections on the relation between meta-analytic methods and theory, and the future of quantitative synthesis methods in educational research.

Throughout, we will place particular emphasis on the importance of using a method that is appropriate for the research question, and how emerging methods allow us to address a broader range of research questions. To facilitate this discussion, it is useful to establish whether the purpose of the analysis is to aggregate or configure the data—or both; we consider this in the next section.

The distinction between aggregating and configuring data

At the broadest level, the overarching purpose of a meta-analysis (and, indeed, any systematic review), is to collate the evidence relevant to a particular topic or research question. Once the evidence is collated, however, the reviewer has a choice of what to do with the studies. The choice is of a synthesis method that can fall anywhere on a spectrum of options, ranging from the *aggregative* to the *configurative*.

Aggregative reviews are defined as those that collect "empirical data to describe and test predefined concepts… The primary research and reviews are adding up (aggregating) and averaging empirical observations to make empirical statements (within predefined conceptual positions)" (Gough, Thomas, & Oliver, 2012, p. 3). Configuration involves placing pieces of evidence in the context of one another; that is, arranging them with the aim of explaining their relationships and interrelationships (Gough, et al., 2012). Most syntheses have elements of both aggregation and configuration. (For a fuller discussion of configuration and aggregation, please see Gough and Thomas's paper in this special issue, "Systematic reviews of research in education: aims, myths and multiple methods".)

Quantitative synthesis, including meta-analysis, can be used for both aggregating and configuring the evidence. For instance, whilst calculating an overall mean effect size estimate is an aggregative process; grouping the studies on the basis of certain characteristics is a configurative enterprise as it involves arranging the data. Thus, a meta-analysis that includes both an overall mean estimate and subgroup analyses has both aggregative and configurative goals.

There has been a shift from the historical quantitative /qualitative divide that focuses on either the type of data or the type of analysis, to a configure/aggregate continuum that

emphasises the goal of the review. Whilst there is still a place for purely aggregative meta-analyses (e.g., to synthesise a set of near-replication interventions, addressing a narrow research question), we are increasingly seeing a need for aggregative-configurative blends. This is because review users are asking increasingly complex questions; not only do they ask 'what works' (the mainstay of classical meta-analysis approaches), they also ask complementary questions such as 'how does it work', 'why does it work', 'for whom', 'under what conditions', and 'at what cost'. In the following sections, we distinguish between quantitative synthesis methods that have a primarily aggregative or primarily configurative goal and consider the types of research questions that they allow us to address.

Purely aggregative statistical meta-analysis

Statistical meta-analysis (see Borenstein, Hedges, Higgins, & Rothstein, 2009; Deeks, Altman, & Bradburn, 2001; Lipsey & Wilson, 2001) in its simplest form is an aggregative approach to synthesis. The findings from individual studies are transformed into a common metric—an effect size—and then pooled into an overall estimate of the effect, with an estimate of its precision (Egger, Davey-Smith, and Phillips, 1997).

In its infancy, most researchers used meta-analysis for pooling the results of related studies to increase power (Egger, Ebrahim, & Smith, 2002). Perhaps because of this simple goal, early meta-analyses were based on a simple statistical model: the fixed-effects model (Hedges & Vevea, 1998). This assumes that all samples are drawn from the same population of studies so that there is only one true (homogeneous) population effect (Cohn & Becker, 2003; Erez, Bloom, & Wells, 1996).

It was not long before reviewers and users of reviews demanded more from meta-analyses. They wanted to extrapolate from the set of collated studies to a broader population of studies. Fixed effect models could not meet this need because the statistical assumptions underlying the models meant that the results could not be generalised to other studies (Raudenbush, 1994). This is because fixed effect models assume that there is no systematic variability between studies (i.e., they are homogeneous). In reality, this is rarely the case, and so models that do not hold that assumption are perhaps more broadly appropriate (Hafdahl, 2007).

To meet this need, random-effects models were proposed. In these models, variability between studies is assumed to reflect both sampling error (within studies) and variability in the population of effects (between studies) (Cooper, Hedges, & Valentine, 2009; Raudenbush, 1994). This makes the findings from random effects model analyses appropriate to generalize to other studies that were not included in the meta-analysis (Lipsey & Wilson, 2001). In practice, the procedure for the random-effects model is similar to fixed-effects models, except that a random error variance component is added to the variance associated with each effect size—it is this feature that takes into account variability in the population effects.

Two factors compelled the subsequent developments in quantitative synthesis. The first was the common experience of meta-analysts that there was variation amongst their studies, both statistical and conceptual, that they wanted to explain. Original meta-analytic methods were simplistic and not suitably designed for dealing with the complexity that is so common in educational research. The second was the increasing diversity of questions being asked by policymakers, decision-makers, and practitioners that were too complicated for existing meta-analytic methods. With the ever-expanding toolkit of primary research designs that catered

for these questions, it was only a matter of time before systematic reviewers also wanted to address these issues. Meta-analysis and other quantitative synthesis methods were developed or adapted to allow for more configurative approaches to synthesis.

Configurative approaches to synthesising studies quantitatively

Sub-group analysis

Whether they are used to explain statistical variation in the effect sizes or to test a priori research questions about how the studies might differ from each other, subgroup analyses (see Lipsey & Wilson, 2001) were one of the first and are still one of the most common configurative approaches to synthesising studies. Subgroup analyses involve grouping the effect sizes by distinguishing characteristics, such as the location of the study or the intensity of an intervention, and are useful for exploring key variables on which we think studies might vary. They are appropriate for categorical variables.

The subgroup analyses are analogous to ANOVAs that are used in statistical analyses in primary research, but adjusted to account for the fact that we are analysing summary statistics rather than raw data (Cooper et al., 2010; Lipsey & Wilson, 2001). The models can be either fixed effect or random effects models, and involve calculating the within-group and between-group heterogeneity. Significant *within-group* heterogeneity indicates that the effect sizes within that subgroup are not statistically similar enough to be grouped together and the group might therefore need to be redefined. Significant *between-group* heterogeneity indicates that the groups of studies are statistically different from each other.

A major limitation of subgroup analyses is that only one variable can be modelled at a time, such as the ‘intensity’ of the intervention. We might also be interested to detect, however,

whether variation in effect sizes can be explained by both the publication date of the study *and* the intensity of the intervention, if we suspect that interventions have become more intense over time. Also, continuous variables need to be converted to categories for inclusion in a subgroup analysis; for example, the continuous variable of publication year would need to be calibrated as categories, such as grouping studies by decade. This can be problematic when natural cut-off points in the data are unclear, forcing the analyst to make arbitrary decisions about category membership that can affect the results and their interpretation.

Meta-regression and multilevel models

Meta-regression (see Lipsey & Wilson, 2001) and multilevel models (see Hox, 2010) allow for multiple variables to be modelled, and can be used with either continuous or categorical variables (that latter need to be converted into dummy variables). They have a similar purpose as subgroup analyses, in that their aim is to explain differences in the effect sizes. In the language of these models, the variables we are testing are called *predictor variables*. Essentially, we are testing whether we can predict the outcome based on what we know about the predictor variable/s in the model. A critical (but often overlooked) output of these models is the unexplained variance, which tells us how much variation in our effect sizes is not explained by our predictor variable/s. If substantial variance remains unexplained, then we need to consider alternative or additional explanatory variables to model.

As with subgroup analyses, meta-regression and multilevel models are adapted from the namesake models used in primary research. Meta-regression can be a fixed or random effects model; a random effects meta-regression is sometimes called a mixed effect model. Mixed effects and multilevel models take into account both sampling error and variability in the population of effects and allow for both within and between-study variation.

Multilevel models view meta-analytic data as being hierarchical, with effect sizes nested within studies (Goldstein, 2003). The differences between multilevel and random effects meta-regression models are beyond the scope of this paper, although it is worth noting that Raudenbush and Bryk (1985; 2002) describe the random-effects model for meta-analysis as a special case of the multilevel regression model. Two benefits of using multilevel analysis instead of random effects meta-regression are flexibility, such as the ability to add further levels to the model (Hox & de Leeuw, 2002), and the range of estimation and testing methods available for multilevel modelling (Hox 2010).

Multilevel models and meta-regression, however, both require more data points (i.e., effect sizes) than subgroup analyses. At least ten studies are recommended for conducting a meta-regression (Higgins & Green, 2011), and a similar number in a multilevel model without predictor variables (e.g., nine studies were analysed in Goldstein, Yang, Omar, Turner, and Thompson, 2000). If predictor variables are included, the number of studies required increases substantially (rules of thumb from multiple regression can be used, such as Tabachnick & Fidell, 2001). Such large numbers of studies are rarely available in most meta-analyses (Thompson & Higgins, 2002). Specifically in education, Ahn, Ames and Myers (2012) reviewed education meta-analyses published between 2000 and 2010 in eight targeted journals. Seventeen (30%) of the 56 meta-analyses that they located had fewer than 30 included studies and would therefore not be recommended to have more than one predictor variable in the regression model, thereby limiting the benefit of this type of modelling over more simple subgroup analyses in these reviews.

Multiple outcomes analysis: Multivariate meta-analysis and robust variance estimation

Historically, meta-analyses considered one outcome at a time. An important statistical assumption of both fixed effects and random effects models is that the outcomes are

independent of each other (Kalaian & Raudenbush, 1996; Lipsey & Wilson, 2001). Studies often report more than one outcome that might be of interest to the meta-analyst, whether it be the same or a similar outcome that has been measured in different ways, or distinct outcomes that the reviewer would like to compare and contrast. In the presence of multiple outcomes that are the same or similar, meta-analysts have used a variety of approaches, including averaging the different outcomes to form a single score; randomly selecting a single outcome from each study for consideration in the meta-analysis; or running separate meta-analyses on the different outcomes. In the case of distinct outcomes, meta-analysts have typically conducted a separate meta-analysis on each outcome, which does not allow direct comparison of the two outcomes.

Whilst these approaches might be appropriate in some situations, they are unsatisfactory when we are explicitly interested in more than one outcome. For example, a meta-analysis—conducted before the advent of multiple outcome meta-analysis methods—of self-concept interventions was forced to focus on global self-esteem as the single outcome (Haney & Durlak, 1998), even though there is strong evidence for the multidimensional nature of the self-concept construct (Marsh & O'Mara, 2008). This potentially underestimates the effects of self-concept interventions given that we see larger effect sizes for self-concept domain outcomes (e.g., academic self-concept, physical self-concept) that were targeted by the intervention compared to those outcomes that were secondary or incidental to the intervention, and global self-esteem was typically not the focus of the intervention (O'Mara, Marsh, Craven, & Debus, 2006).

Methods have been developed to allow for multiple outcomes in the same meta-analytic model. Analogous to the difference between MANOVA and ANOVA analyses in primary

research, running multiple (ANOVA) analyses on the separate outcomes can increase the risk of Type I error, whereas including them in the same model (as in MANOVA) allows us to test the effects of the independent variables on the outcomes simultaneously—reducing the risk of Type I error. Beyond the statistical benefits, having multiple outcomes in the same model allows us to test a broader range of questions. For instance, does an intervention have equal benefits across a range of outcomes, or do some outcomes improve more than others?

The two main approaches to multiple outcome meta-analysis are multivariate multilevel modelling (Hox, 2010; Raudenbush & Bryk, 2002) and the robust variance estimation method (Hedges, Tipton, & Johnson, 2010). Note that both of the following methods can be used in other instances of dependencies in the data, such as measurement at multiple time points of the same person, or multiple studies conducted by the same researchers.

In the multivariate multilevel model, it is assumed that effect sizes within the same study are more similar (more strongly correlated) than effect sizes in other studies, and so within-study and between-study correlations are modelled (see Hox, 2010; Kalaian & Kasim, 2008a, 2008b; Kalaian & Raudenbush, 1996; O'Mara, 2009; Riley, Thompson, & Abrams, 2008). By accounting for the nested structure in multilevel modelling and by modelling the effect size dependencies, the problems with dependencies within the data are minimised. Multilevel model meta-analyses are surprisingly uncommon, despite being proposed thirty years ago (Raudenbush, 1994; Raudenbush and Bryk, 1985). The key limiting factor is that the correlation between outcomes needs to be known (or reasonably imputed); unfortunately, primary studies often do not report the correlations between outcomes. Without the known correlation from each study, the improved precision of estimates in the multivariate

multilevel model compared to a random effects model are likely to be negligible (O'Mara, 2009).

The robust variance estimation method is an adjustment to the meta-regression model to take into account dependent effect size estimates (Hedges et al., 2010). The computation of the variance estimate is adjusted. Importantly, this model does not require knowledge of the covariance structure of the dependent estimates (Hedges et al., 2010), and so is not limited when primary studies do not report the between-study correlations.

Neither of the above methods are particularly well-represented in the meta-analytic literature. Whilst a lack of off-the-shelf software is partly to blame, the conceptual complexity of the models could also be a deterrent.

Analysis of indirect comparisons: Network meta-analysis and ranking systems

There are two key reasons why the next meta-analytic methods that we consider developed: the need for decision-makers to know the comparative effectiveness of different intervention options, and the challenge of having a multitude of related interventions and comparators that have been evaluated in different combinations.

Having a range of approaches to choose from is a typical decision-making situation, and many policy-makers and practitioners want to know about the comparative effectiveness of different interventions. In other words, of a range of possible interventions, which is most likely to work? This question shifts the emphasis of traditional meta-analyses from 'does it work' to 'which works better/best'. Fortunately, relatively new methods of network meta-analysis (see Caldwell, 2014; also known as indirect comparisons and multiple-treatments meta-analysis) have been developed in response to this need.

Network meta-analysis allows us to consider the comparative effectiveness of interventions that may or may not have been evaluated directly against each other, and thereby truly benefits from the pooling of knowledge across studies. Network meta-analysis considers the set of studies as a network of direct comparisons (i.e., the comparisons made within a primary study) and indirect comparisons (i.e., those made through one or more common comparators) (Mills et al. 2013). A ranking of intervention effectiveness can be produced.

A further benefit of network meta-analysis lies in the way that it allows indirect comparisons. It is not uncommon, once all of the studies have been assembled in a meta-analysis, to observe a patchwork of interventions and comparators. For instance, some studies might compare the intervention to a no-treatment condition (i.e., the control group continues as usual), others might compare to an 'attention placebo' (i.e., the control group is provided with a low-intensity alternative program that is designed to account for the possible interest or excitement induced by participation in a new program), and others still will compare with a wait-list condition. The situation gets even more complicated when different interventions are offered, whether they be variations of a similar intervention (e.g., peer tutoring versus peer tutoring with study training) or completely different interventions (e.g., peer tutoring versus feedback). Traditional meta-analytic methods struggle to deal with these situations. Commonly (and we would argue, inappropriately), effect estimates are combined and the heterogeneity is ignored. In other situations, subgroup analyses are conducted to determine whether the differences in the intervention or comparison condition are associated with differences in the observed effect. In some cases (that can be either sensible or not, depending on the research question being asked), the review's inclusion criteria can be refined to ensure a more homogenous set of studies (although this runs the risk of throwing away potentially

valuable and informative data). Network meta-analysis means that we do not have to throw data away or make conceptual compromises, as it allows us to model both direct and indirect comparisons.

Network meta-analysis is not suited to all datasets, however. A problem in terms of the power and reliability of the analyses can occur where there is little evidence available on a direct comparison or where there is a marked imbalance in the amount of evidence for each intervention (Thorlund & Mills, 2012).

The increase in popularity of network meta-analysis and other ranking approaches can be seen as an increased awareness on the part of the systematic review community that individual comparisons can only take us so far in terms of informing decisions. This realisation also motivated the following approaches, which include other analysis types to complement the statistical meta-analysis.

Multi-method approaches to synthesising evidence

Mixed methods synthesis

A mixed methods *review* is a type of review that combines different types of data, often by utilising multiple methods of analysis. Mixed methods reviews are not necessarily reviews of mixed methods primary studies—the design of the review is distinct from the design of the included studies. In the case where multiple analytical tools have been used, a mixed methods review has employed a mixed methods *synthesis* approach. Mixed methods synthesis (see Harden & Thomas, 2005; 2010) can take many forms. Harden (2010, p.4) states that there are three ways in which reviews can be mixed:

1. The types of studies included in the review are mixed; hence, the types of findings to be synthesized are mixed.
2. The synthesis methods used in the review are mixed—statistical meta-analysis and qualitative [although we note that other types of evidence can also be used, such as theoretical papers or economic evaluations].
3. The review uses two modes of analysis—theory building and theory testing.

Given the focus in this paper on meta-analysis, we specifically refer here to mixed methods syntheses that include a quantitative synthesis component, although it is noteworthy that a mixed methods synthesis does not need to have a quantitative analysis component. A typical example (Thomas, Harden, Oakley, Oliver, Sutcliffe, et al., 2004) is a review in which qualitative data consisting of views of children about healthy eating were then mapped against the interventions (quantitative data) to see whether interventions that addressed the views of the children were more or less effective than those that did not. In another example, the results of a synthesis of theoretical papers were used to explore theories of change in interventions of community engagement in public health interventions (O'Mara-Eves, Brunton, McDaid, Oliver, Kavanagh, et al., 2013).

In both of these cases, and in most mixed methods syntheses, the different evidence types are usually processed separately first, and then combined in some way. How they are combined is up to the creativity of the research team and the requirements of the research question (see Sandelowski, Voils, & Barroso, 2006, for different designs of mixed research synthesis).

Critical to understanding why and when we should use these approaches is a consideration of the epistemological issues that these reviews address. Whilst traditional meta-analyses focus

on evidence that explores whether an intervention works, other types of evidence can allow us to integrate an understanding of why something (might) work (Harden & Thomas, 2005). Furthermore, combining different types of evidence can allow us to blend micro (individual) and macro (collective) perspectives, which reflects the real-world complexity of the situations that we are investigating.

Such reviews can offer several advantages over a meta-analysis on its own, including:

- The opportunity to consider barriers and facilitators to intervention success that might only be elicited through qualitative primary research.
- The possibility to explore how contextual differences might translate into differential effects.
- A deeper understanding of the theories and mechanisms of change that are being proposed and tested in the intervention might be gathered.
- The selection of moderators and predictor variables for inclusion in the meta-analysis can be better informed, can avoid data dredging, and can examine the issues that are of most importance to the review's audience and populations of interest.
- The review can have enhanced relevance because it can take a holistic approach to exploring what works, why it works, for whom, and at what cost.

Using large-scale datasets in conjunction with meta-analysis

When we talk about large-scale datasets, we are referring to datasets with many participants and can include routinely collected data (e.g., census data, National Pupil Database), longitudinal cohort and panel data (e.g., Longitudinal Study of Young People in England, the UK Millennium Cohort Study), and detailed cross-sectional survey data (e.g., Crime Survey for England and Wales, the multi-national Programme for International Student Assessment).

There are several ways in which this type of data can be used in conjunction with meta-analysis.

Firstly, the data can be used as raw data in a meta-analysis. Whilst meta-analyses traditionally have used the summary statistics (e.g., means and standard deviations) reported in primary study reports, there is an increasing move for so-called 'individual participant data (IPD) meta-analysis' (see Riley, Lambert, & Abo-Zaid, 2010), which uses the data collected from the individual participants in the studies. Methods for IPD meta-analysis are rapidly evolving (see the Cochrane IPD meta-analysis methods group website, http://ipdmamg.cochrane.org/methods-research), but the basic approach is generally either one or two step (Debray, Moons, Abo-Zaid, Koffijberg, & Riley, 2013). In a one-step approach, the IPD from all studies are modelled and adjusted for the clustering of participants within studies. In a two-step approach, the IPD are first analysed separately for each study, then the results of these separate analyses are synthesised using standard meta-analytic techniques (e.g., fixed, random, or multilevel models). There are many potential advantages of IPD meta-analysis (Riley et al., 2010), but perhaps the key advantage is that the meta-analyst can select from the full dataset the data (participants, moderator/predictor variables, outcomes) that are of interest, rather than relying on the study authors' reporting of this data (which might not have the same aim as the review's focus).

There is exciting meta-analytic work in education that is capitalising on rich datasets with individual student data. A particular highlight in the UK is the work being done through the Education Endowment Foundation (EEF). The EEF commissions primary evaluations of interventions aimed at addressing disadvantage in school settings, and is in the process of building a database of projects and intervention data that has been specifically designed to be

matched with national test results. Already, some of the data from completed evaluations has been meta-analysed alongside other relevant studies in the EEF Toolkit (see https://educationendowmentfoundation.org.uk/toolkit/). It is anticipated that by 2027, the EEF database will contain data from about 500 interventions including over 1 million pupils, and has large potential for meta-analyses to identify comparative impact and assess variation in impact (Higgins, 2014). The database also includes results from non-EEF interventions and systematic reviews. This inclusion lends strength in terms of providing more information, but also introduces conceptual heterogeneity that is not explored or controlled for in the toolkit. In particular, the approach largely ignores any differences due to the type of comparator used. Whilst this is not necessarily problematic in trying to establish whether an intervention 'generally works', it is problematic for ranking because the intensity or effectiveness of the comparator could lead to smaller or larger effect size estimates. For example, an intervention might undeservedly be ranked higher simply because it is compared against a weaker comparator than another intervention. A related issue is that, whilst the EEF interventions have been designed to be matched with (UK) national test results, the non-EEF interventions have not. Such concerns make it difficult to assess how applicable the findings are to a particular context (i.e., "will the intervention work at my school with my pupils?"), which partly undermines the purpose of the database as a practitioner-focused decision-making tool.

Secondly, large-scale data can be analysed and the results reported alongside the meta-analysis (see Marsh, Bornmann, Mutz, Daniel, & O'Mara, 2009). There are strengths and weaknesses associated with both meta-analyses and analysis of large datasets. Primary studies can reduce threats to internal validity, but a single study cannot evaluate the generalizability of the results. A meta-analysis can explore consistency and generalizability of the results across different studies, but the typical lack of access to the raw data means that

meta-analysts can usually only explore variables that are reported in the original studies—this is particularly a problem when there is significant statistical heterogeneity across the studies (and one of the main reasons why IPD meta-analyses are gaining popularity). Meta-analyses and secondary data analysis can therefore be complementary when IPD meta-analysis is not possible or appropriate. As an example, Marsh et al. (2009) juxtaposed the results of a large primary study (around 10,000 data points) looking at the potential for gender bias in the academic peer review process, with a multilevel meta-analysis of gender bias studies. Both analyses found no consistent evidence of a gender bias, which was in contrast to an existing meta-analysis of peer review data that found a small gender bias in favour of males (Bornmann, Lutz, & Daniel, 2007). In this way, the combined strength of the secondary data analysis with the new meta-analysis were able to overturn the conclusions of the original review (Marsh & Bornmann, 2009).

Other ways in which large-scale data can be used with meta-analysis, but have not yet been applied or do not have established methods, include:

- To identify variables to explore in the meta-analysis, in a similar way that was proposed for qualitative evidence in the section on mixed methods synthesis.
- To explore factors that could not be examined in the meta-analysis, because the factors were not reported in the primary studies.
- To explore issues of generalisability and transferability. Although the generalisability of statistical findings requires a well-defined population to which a researcher intends to extrapolate the findings (Hedges, 2013; O'Muircheartaigh & Hedges, 2014), the inference population for a typical meta-analysis is usually unclear because of heterogeneity in the dataset. As one possible way of dealing with this, O'Muircheartaigh and Hedges describe methods for estimating the generalisability

from an experiment to different populations by exploring covariates that could be adapted for meta-analytic purposes.

These last three ways of using large-scale data with meta-analysis are currently being investigated by the authors and their colleagues.

Qualitative Comparative Analysis (QCA)

More recently, the potential of using qualitative comparative analysis for synthesising research studies has been explored (Thomas, O'Mara-Eves, & Brunton, 2014; Brunton, O'Mara-Eves, & Thomas, 2014). Despite its name, it is not strictly a qualitative approach: the outputs are numerical and effect sizes are used to determine whether a study belongs to the set of effective or the set of not-effective interventions (or a fuzzy set of 'partially effective interventions'). QCA differs from the methods discussed so far as it has no basis in inferential statistics or null hypothesis significance testing. It is a method used to explore complex causality by investigating which combinations of particular conditions evident in the studies are more often found in effective (and non-effective) interventions (Ragin, 2008a). For instance, a public health commissioner might be interested in whether offering incentives is a necessary condition for intervention effectiveness, regardless of the rest of the content of the intervention.

This is a critical difference to the logic behind the methods discussed so far. In QCA, we can model multiple causal pathways to the same outcome (Thomas et al., 2014). The conditions (characteristics of the studies and interventions) can be necessary for intervention success (the condition must be present for the desired outcome to be achieved, regardless of whatever else is done), sufficient for intervention success (this is all that is needed to achieve the desired outcome), or not relevant to the success of the intervention (it does not matter if this is present or not). In contrast,

> statistical methods that are based on the exploration and explanation of correlations between variables are sometimes ill-suited in the analysis of causal pathways… because correlational approaches test simultaneously for the 'success' and 'failure' of covariates, they are unable to identify the importance of the component 'x' in intervention A when it is also present in unsuccessful intervention B. (Testing for interaction is usually impossible in systematic reviews because of a lack of data.) (Thomas et al., 2014, p. 68).

QCA is relatively new to the systematic review synthesis methods toolbox, but shows great promise in being able to deal with complex interventions and unpick the key characteristics to intervention effectiveness (Thomas et al., 2014). Moreover, it can be used with a small number of studies, is computationally easy, and has free and easy-to-use software (Ragin, 2008b). Perhaps the main disadvantage is the metrics: the operational definitions of coverage and consistency are tricky to grasp at first, and might be difficult to communicate to new audiences (see Brunton et al., 2014, for an application of the method).

All of the multi-method approaches to synthesis have developed or been adapted from primary research methods to enable us to address the increasingly complex informational needs of review users. Particularly in educational research, we need methods such as QCA because we often do not have the data to model every variable in a reliable way, and we need mixed methods approaches to help us to understand different perspectives and to take account of how unobservable phenomena impact on what we can observe.

Challenges for education meta-analyses

This section focuses on two issues that are particular challenges for meta-analyses of educational research: publication bias and interpreting effect sizes.

Publication bias

Concerns about publication bias (see Sutton, 2009) have been raised after observations that research evaluations showing beneficial and/or statistically significant findings are more likely to be published than those that have undesirable outcomes or non-significant findings; basing a review on biased data could lead to biased findings (Higgins & Green, 2011). There is little doubt that publication bias can be a problem in educational research. Lipsey and Wilson's (1993) synthesis of educational and psychological meta-analyses, for example, found that published studies had mean effect sizes that were 0.14 standard deviations larger than the mean effect sizes of unpublished studies, across 92 meta-analyses that provided separate effect sizes for the two types of studies. We would not expect to see a systematic difference if there was no publication bias.

This is a particular challenge for education meta-analyses because it is an issue that is underexplored in educational reviews. Ahn et al. (2012) found that 57% of 56 education meta-analyses that they reviewed did not test/report whether publication bias might be a concern, which suggests that many education meta-analysts are not fully reporting risks to the strengths of their conclusions.

Unfortunately, it is difficult to assess publication bias because we cannot know the extent of what has not been published. Nevertheless, technical approaches to detecting possible bias have been developed, including:

- *Funnel plots* (such as Egger, Smith, Schneider, Minder, 1997) are the most popular approach for detecting publication bias, but Torgerson (2006) warned against using funnel plots in fewer than 20 studies, which would make it unusable for many meta-analyses. Moreover, their interpretation is subjective and can be misleading (Lau et al., 2006). Tests for funnel plot asymmetry, which attempt to provide an objective interpretation of the plots, are only appropriate in a minority of meta-analyses (Ioannidis & Trikalinos, 2007):
    - Begg's test (Begg & Mazumda, 1994) starts with a funnel plot and then tests the asymmetry, but is not recommended by the Cochrane Collaboration[2] due to low power (Higgins & Green, 2011).
    - Eggar's test (Egger, et al., 1997) is recommended for continuous outcomes with mean differences effect sizes (Higgins & Green, 2011).
    - Several tests of funnel plot asymmetry are available for odds ratio dichotomous outcomes, but each has pros and cons, while other dichotomous outcomes have no recommended test; see Higgins and Green (2011) for a review.
- *Rosenthal's (1979) Fail-safe n*, whilst traditionally a popular approach, is now generally regarded as inappropriate (Sutton, Song, Gilbody & Abrams, 2000).
- *Modelling of publication bias* can take many forms. For instance, subgroup analyses or meta-regressions can be conducted, with published and unpublished studies as the groups/predictors (Torgerson, 2006). A significant difference would indicate that effect sizes are systematically different for the two types of literature (although note that this does not necessarily indicate the presence of publication bias). Other

[2] The Cochrane Collaboration (http://www.cochrane.org/) is the health equivalent of the Campbell Collaboration in the social sciences. The Cochrane Collaboration is a non-profit network of researchers conducting systematic reviews in healthcare, with method experts that develop guidance on how to conduct high quality systematic reviews.

modelling approaches include selection models (which are not widely used due to their complexity; see Sutton et al., 2000 for an evaluation) and the precision effect test (a meta-regression based approach that is increasingly popular in economic meta-analyses but requires further evaluation; see Stanley, 2008 for a description).

- *Trim and fill* (Duval & Tweedie, 2000a, 2000b) calculates the effect of potential data censoring on the outcome of the meta-analyses; it is a nonparametric, iterative technique that examines the symmetry of effect sizes plotted by the inverse of the standard error and then imputes the values of 'missing' studies. A test is then applied to determine whether the imputed values would substantially affect the overall mean effect size. Whilst perhaps more promising than many other approaches, it still suffers in that any observed asymmetry is not necessarily due to publication bias (Sutton et al., 2000).

It might not be (currently) possible to satisfactorily detect or adjust for publication bias, but it is something that meta-analysts should consider. A central tenet of this paper is that meta-analyses (and systematic reviews more broadly) need to be useful and usable for their intended audience; part of this involves the responsibility of the reviewer to be transparent about threats to the robustness of the review findings, including the potential for publication bias. As such, even if none of the abovementioned tests are used, meta-analysis authors should alert the reader to the possibility that this bias might be present.

We further suggest that mixed methods approaches may help determine the robustness of the quantitative findings where the aforementioned technical approaches fail, as qualitative evidence should not be vulnerable to the same causes of publication bias (i.e., different issues threaten the likelihood of qualitative evidence being published). We are only recently seeing the potential for qualitative research to provide evidence on intervention impact in systematic

reviews (Petticrew, 2015), and this evidence could sit alongside the meta-analytic findings to help to consolidate the strength of the conclusions.

Interpreting the results

There is a slow change in the way that meta-analysts are interpreting effect sizes. Presenting meta-analytic findings is a particular challenge for education because practitioners (teachers and educators) and educational decision-makers are becoming increasingly evidence-hungry, but need information that is digestible for non-statisticians. The popularity of resources such as the US Institute of Education Sciences *What Works Clearinghouse*, Hattie's (2008) *Visible Learning*, and the EEF's (2014) *Teaching and Learning Toolkit* are indicators that there is a demand for accessible summaries of the evidence base.

Historically, most meta-analyses would reference Cohen's (1977) categories of 'small', 'medium', and 'large' effect sizes, which roughly correspond to *d*'s of .2, .5, and .8 or *r*'s of .1, .25, and .4, respectively. Even now, this is a commonly cited interpretation of the magnitude of effect estimates. There are, however, at least three reasons why there is a shift away from this way of describing effect sizes:

1. The labels assigned by Cohen (1977) were not empirically or mathematically derived. Cohen (1992) acknowledged: "Although the definitions were made subjectively, with some early minor adjustments, these conventions have been fixed since the 1977 edition of SPABS [Cohen's book, *Statistic Power Analysis for the Behavioral Sciences*] and have come into general use." These categories have had little empirical testing. Lipsey and Wilson (2001) compared the Cohen (1977) values with the quartiles of the distribution of mean effect sizes from over 300 meta-analyses of psychological, behavioural, or educational interventions (reported in Lipsey & Wilson, 1993). The bottom quartile was effect size ≤ .3,

median effect size = .5, and top quartile was effect size ≥ .67. Although these are not far from Cohen's values, they could certainly represent real-world differences.

2. The labels ignore context. As more and more meta-analyses are conducted, it is increasingly clear that mean effect sizes can vary dramatically from topic to topic. The distribution of effect sizes in the 300-plus meta-analyses in Lipsey and Wilson's (1993) and in the 800-plus meta-analyses in Hattie's (2008) meta-meta-analyses highlight this issue. There is therefore an increasing call for the magnitude of effect sizes to be assessed in the context of the topic area, rather than trying to generalise magnitude across incredibly diverse topics (Lipsey et al., 2012). For example, in some topic areas, a 'small' effect size of .10 might represent an important improvement on previous effect estimates, or might be the best that we can expect for the given outcome.
3. The labels ignore practical (real-world) significance. Whilst an effect size of .10 might look statistically small, that .10 of a standard deviation might actually represent a large or important real-world effect. The users of meta-analytic—policymakers, decision-makers, practitioners—are increasingly wanting to know what is the practical significance of a given intervention (e.g., how many more students will pass exams, etc).

Another commonly-reported way of interpreting effect sizes is to rely on the statistical significance of the effect estimate. However, this is also flawed, as Lipsey et al. (2102) argue:

> The *p*-values characterize only statistical significance, which bears no necessary relationship to practical significance or even to the statistical magnitude of the effect. Statistical significance is a function of the

> magnitude of the difference between the means, to be sure, but it is also heavily influenced by the sample size, the within samples variance on the outcome variable, the covariates included in the analysis, and the type of statistical test applied. (p. 3)

More informative interpretations are gradually becoming more common. Some promising directions are:

- Converting back to the original metric (see Lipsey & Wilson, 2001)—this is arguably the best way to interpret the effect size as it allows the reader to see the real-world impact of the intervention (e.g., actual improvement in test scores or grade points);
- Visual representations, such as Hattie's (2008) 'barometers' and the EEF (2014) Toolkit's cost and evidence security estimate symbols, although these simplify effect sizes in ways that could be misused or be unintentionally misleading and therefore must be presented with caution;
- EEF's (2014) 'average impact', which is defined as "the additional months' progress you might expect pupils to make as a result of an approach being used in school, taking average pupil progress over a year is as a benchmark." (p. 4), although this measure ignores differences in progress for different ages and types of test;
- Number needed to treat, which is appropriate for binary variables and is loosely defined as the number of people that would need to receive the intervention for one person to benefit (see Citrome, 2007, although note criticisms e.g., Hutton, 2000); and
- Cost-benefits and cost-effectiveness, which could be useful to decision-makers and commissioners, but rely on accurate cost data (see Lipsey et al., 2012, for a brief introduction to using cost data in interpreting meta-analytic effect sizes).

More work needs to be done to develop clear ways for meta-analysts to report the results of their meta-analyses in ways that are meaningful for their audience. Whilst other topics covered in this paper focus on making meta-analyses fit for purpose, improving the interpretation of effect sizes is an attempt to make meta-analyses fit for use. Hopefully this will be a direction for future methodological work.

Reflections on the relation between meta-analysis and theory

When thought of as a purely aggregative process—as classical meta-analyses have historically tended to do—meta-analysis is more suited to theory testing than theory generation. There is reasonable scope for theory generation in a purely aggregative meta-analysis if main effects (i.e., the effect of one independent variable on a dependent variable, ignoring any other potential moderators or mediators) are being proposed. This is, of course, rather limiting. As more configurative approaches to meta-analysis evolve and new methods emerge, so too does the scope for using meta-analysis as a theory generation tool. There are three reasons why meta-analysis is becoming an increasingly useful tool for theory generation.

Firstly, configurative approaches to meta-analysis allow us to explore contingencies, rather than just main effects. By being able to model moderator variables (subgroup analyses), predictor variables (meta-regression), multiple 'independent variables' – the intervention (network meta-analysis), multiple outcomes (multivariate meta-analysis), and multiple pathways to effectiveness (combining meta-analysis with QCA), we can now develop theories that more accurately reflect the complexity of real world situations. This, of course, is limited by the extent to which the included studies report the variables and data needed to generate the theories.

As an example of how configurative meta-analysis can lead to theory generation, we consider academic feedback, which is a commonly used method to enhance achievement and self-esteem in the classroom. Prior to a 2007 meta-analysis by Hattie and Timperley, feedback had been associated with both positive and negative consequences. Their meta-analysis found that feedback can be effective for enhancing performance, but can also have negative effects if directed at the personal level (usually in the form of non-contingent praise). From these findings, the authors and subsequent researchers could generate new theories about why feedback sometimes worked and sometimes did not.

Secondly, emerging methods in meta-analysis that emphasise individual participant data and disaggregated data are enabling greater creativity for theory building, as we are less constrained by the primary study authors' choices about what to analyse and report. This is particularly important for fields of research that are evolving, so that issues that were the zeitgeist when the original study was produced might no longer be the issues of interest. Whilst this helps us to circumvent the issue of reporting, however, we are still limited by the data collected in the primary studies (and our ability to gain access to it).

Finally, we are re-understanding the issue of heterogeneity in meta-analysis. Researchers have long criticised meta-analyses that mix 'apples and oranges' (Eysenck, 1994; Wolf, 1986; see Cortina, 2003), and heterogeneity was often seen as a problem (e.g., of small samples, unreliable measurement, and other artefacts) that need to be "corrected" (e.g., Hunter & Schmidt, 2004). Whilst there are obviously limits to the extent to which it is sensible to combine vastly different studies, and there are clearly methodological issues that need to be modelled or "corrected", there is an increasing realisation that variability observed

in our datasets could also be due to meaningful differences. Indeed, Glass (2000), the father of meta-analysis, stated:

> Of course it [meta-analysis] mixes apples and oranges; in the study of fruit nothing else is sensible; comparing apples and oranges is the only endeavor worthy of true scientists; comparing apples to apples is trivial

This shift from problematizing variability to trying to explain it was necessary for meta-analysis to be useful as a theory generation tool.

Quantitative synthesis, however, has always been about more than just theory testing and theory generation. They have long been used to describe the current evidence base, to highlight gaps, and to identify areas for future research. Systematic research synthesis methods have allowed us to reflect the state of current knowledge—incomplete and potentially mistaken though it may be. Explicitly or implicitly, they have made it clear that, in many areas, we are unlikely to have the levels of evidence available to address many questions (generally because of practical or ethical reasons). In a society where many believe in evidence informed policymaking and that "statements without evidence are just opinions" (quote from John Hattie in an interview with Evans, 2012), however, we need to make the best of what evidence we do have. Dismissing incomplete or low quality evidence bases outright is wasteful; lessons can still be learnt—even if the lesson is not to do the same thing again! The skill and the responsibility of research synthesists lie in being able to distil the useful elements from the compound evidence base in an unbiased and meaningful way.

We are increasingly realising that the hypothetico-deductive approach is less suited to some of the more fragmented research areas that are typical in educational research and are re-

evaluating how synthesis methods meet the needs of the evidence-informed movement. Petticrew (2015) argued that it is often more useful to phrase review research questions in terms of 'what happens' rather than 'what works':

> …when one moves to reviews of more complex, socially embedded interventions, hypothesis-testing is not only difficult, it may not even be desirable. This is because evaluating complex social interventions purely in terms of whether they 'work' or 'do not work' can be simplistic and misleading. Instead, systematic reviews in these circumstances… should aim to assemble a range of examples of what happened when that intervention was implemented in different contexts. (p. 2)

In summary, meta-analysis specifically—and quantitative synthesis more broadly—can be used for theory testing, theory generation, and reflecting the current state of the evidence base. New approaches that integrate different types of evidence (such as mixed methods synthesis and the use of large-scale datasets) or that apply different logic to synthesising the evidence (such as QCA) have been developed to support these different purposes and the needs that they are inherently addressing.

Moving forward

Meta-analysis is a constantly evolving method. Importantly, it is evolving not just in education, but in many disciplines. We suggest that it is necessary to keep pace with the developments in meta-analysis in other disciplines—particularly health, in which there is a lot of investment in synthesis methods development, where the value of understanding that complexity cannot always be modelled and that more 'systems perspectives' have a role to

play (Petticrew, 2015), are beginning to be appreciated. However, not all approaches will easily transplant, and some issues that are unique or particularly important to education will need to be addressed from within the educational research community; for example, multilevel analysis approaches largely emanated from educational researchers, because the issue of nested designs is particularly pertinent to educational research.

Whilst it might be difficult to keep abreast of all of the latest developments across many disciplines, it is worthwhile regularly scanning methods journals such as *Research Synthesis Methods* and *Systematic Reviews*, as well as journal special issues such as this one, and the methods publications of the Campbell and Cochrane Collaborations, to locate the more important developments that are likely to become widely adopted.

A further recommendation is to not get too bogged down in statistical tweaks (of which there are many – especially for publication bias and other adjustments for artefacts). In most cases, improvements in precision of the effect estimate are so small as to not be worth the 'black box' of methods that we end up with. It is more important to focus on the bigger issues, such as, do the assumptions of the statistical test and the hypothesis that it is implicitly testing allow me to answer my research question? It is all too easy to forget the actual hypothesis being tested by a particular analysis, which is evident in the way that the 'findings' of the result are often misworded. Once the big issues have been dealt with, some key issues within the particular domain of research (e.g., clustering, reliability of the scales used, small *n* in cells in a subgroup analysis) should be considered. In other words, it is not necessary to apply every adjustment and every new exciting method in every review.

When planning a meta-analysis, it is useful to first consult the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, see http://www.prisma-statement.org/). Although it is designed for review authors to ensure that they have reported their meta-analyses appropriately, it serves as a good reminder of the key features of the method that need to planned and recorded, such as how risk of bias will be treated, both within the primary studies and across the studies.

Conclusions

Meta-analysis is an ever-growing field of related methods. Over the four decades since its formalisation, we have changed the way that we think about evidence, with a shift from the historical quantitative /qualitative divide that focuses on either data or analysis, to a configure/aggregate continuum that emphasises the goal of the review. Through this shift, and by de-problematizing observed heterogeneity, we have developed a stunning array of methods that allow us to test a broader range of research questions, including those that can assess multiple outcomes (multivariate approaches) and compare multiple interventions (network meta-analysis). Innovative new approaches that combine with other analysis techniques (e.g., mixed methods synthesis, QCA, and secondary data analysis) add to a toolkit that enables increasingly fit-for-purpose synthesis. We are also now gathering more raw data and relying less on just the published findings (as in IPD meta-analysis) so that we can analyse what is relevant to the review, not just what the primary study authors were interested in.

Overall, meta-analysis is now more suited to theory generation than in its early days. Moreover, it now places a greater focus on the relevance of the review and its findings to the intended audience. The focus of a meta-analysis should always be on whether the data and

the analysis method chosen allow you to address your research question, and this research question should accurately reflect the questions being asked by the intended users of the review.

References

Ahn, S, Ames, AJ, Myers, ND. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research, 82*, 436-476.

Borenstein, M., Hedges, L., Higgins, J. and Rothstein, H. (2009) *Introduction to meta analysis*. Chichester: Wiley and Sons

Caldwell, DM (2014). An overview of conducting systematic reviews with network meta-analysis. *Systematic Reviews, 3*, 109.

Citrome L. (2007). Show me the evidence: Using number needed to treat. *Southern Medical Journal*, *100*(9), 881-884.

Begg CB, Mazumdar M. (1994) Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088-1101.

Bornmann, L., & Lutz, R., & Daniel, H. D. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics, 1*(3) 226–238.

Brunton G, O'Mara-Eves A, Thomas J. (2014). The 'active ingredients' for successful community engagement with disadvantaged expectant and new mothers: a qualitative comparative analysis. *Journal of Advanced Nursing, 70*(12), 2847-2860

Cohen, J. (1977) *Statistical power analysis for the behavioral sciences.* NY: Academic Press.

Cohen J (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*, 243–253.

Cooper HM, Hedges LV, & Valentine JC. (eds). (2009). *The Handbook of research synthesis and meta-analysis (2nd edition).* New York: Russell Sage Foundation.

Cortina, J. M. (2003). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods, 6,* 415-439.

Debray TP, Moons KG, Abo-Zaid GM, Koffijberg H, Riley RD. (2013). Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? *PLoS One, 8*(4):e60650. doi: 10.1371/journal.pone.0060650.

Deeks, J. J., Altman, D. G. and Bradburn, M. J. (2001) Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In M. Egger, G. D. Smith and D. G. Altman (Eds.), *Systematic Reviews in Health Care: Meta-Analysis in Context*, Second Edition. BMJ Publishing Group, London, UK.

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89-98.

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

Education Endowment Foundation (2014). *Sutton Trust – EEF Teaching and Learning Toolkit* (version September 2014). London: Education Endowment Foundation.

Egger, M., Davey-Smith, G. and Phillips, A.N. (1997) Meta-analysis: principles and procedures. *BMJ, 315*, 1533-1537.

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

Egger, M., Ebrahim, S., & Smith, G. D. (2002). Editorial: Where now for meta-analysis? *International Journal of Epidemiology, 31*, 1-5.

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology, 49*, 275-306.

Evans, D. (2012). “He's not the messiah…” Interview with John Hattie. *Times Educational Supplement Magazine,* 14 September 2012. Accessed at https://www.tes.co.uk/article.aspx?storycode=6290240 on 17 August 2015.

Eysenck, H. J. (1994). Systematic Reviews: Meta-analysis and its problems. *British Medical Journal, 309,* 789-792.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5,* 3-8.

Glass G. V. (2015). Meta-analysis at middle age: a personal history, *Research Synthesis Methods,* doi: 10.1002/jrsm.1133.

Goldstein, H. (2003). *Multilevel statistical models* (3rd Rev. Ed). London: Hodder Arnold.

Goldstein, H., Yang, M., Omar, R., Turner, R. and Thompson, S. (2000), Meta-analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 49,* 399–412. doi: 10.1111/1467-9876.00200

Gough D, Thomas J, & Oliver S (2012). Clarifying differences between review designs and methods. *Systematic Reviews, 1*(28).

Haney, P., & Durlak, J.A. (1998). Changing self-esteem in children and adolescents: A meta-analytic review. *Journal of Clinical Child Psychology, 27*(4), 423-433.

Harden A. (2010) *Mixed-methods systematic reviews: integrating quantitative and qualitative findings* [monograph on the internet]. London, UK: FOCUS: Technical Brief no. 25; 2010 [accessed 10 July 2015]. Available from: www.ncddr.org/kt/products/focus/focus25/

Harden A, Thomas J. (2005) Methodological issues in combining diverse study types in systematic reviews. *Internal Journal of Social Research Methodology, 8*, 257–271.

Harden A, Thomas J. (2010). Mixed methods and systematic reviews: Examples and emerging issues. In: Tashakkori A, Teddlie C, editors. *Sage handbook of mixed methods in social & behavioral research*. 2. Thousand Oaks, CA: SAGE; pp. 749–774.

Hafdahl, A. (2007). Combining correlation matrices: Simulation analysis of improved fixed-effects methods. *Journal of Educational and Behavioral Statistics, 32*, 180–205.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Abingdon: Routledge.

Hedges, L.V., Tipton, E., and Johnson, M.C. (2010) Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39-65. Erratum in 1(2): 164-165.DOI: 10.1002/jrsm.5

Hedges LV (2013) Recommendations for practice: Justifying claims of generalizability. *Educational Psychology Review, 25*, 331-337.

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

Higgins J, Green S (editors). (2011) *Cochrane handbook for systematic reviews of interventions version 5.1.0* (updated March 2011). URL: www.cochrane-handbook.org (accessed 15 June 2015).

Higgins, S. (2014). *Meta-analysis, education research & the Sutton Trust-EEF Teaching and Learning Toolkit.* Presented at CEM's 30 Years of Evidence in Education conference, 23 September 2014, London, UK.

Hox, JJ. (2010). *Multilevel analysis: Techniques and applications, second edition (Quantitative Methodology Series).* Hove, GB: Routledge.

Hox, J. J. & de Leeuw, E. D. (2003). Multilevel models for meta-analysis. In S. P. Reise & N. Duan (Eds.). *Multilevel modelling: Methodological advances, issues, and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

Hutton JL. (2000). Number needed to treat: properties and problem. *Journal of the Royal Statistical Society Series A, 163*(3), 403-419.

Ioannidis JP, Trikalinos TA. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Canadian Medical Association Journal, 176,* 1091-1096.

Kalaian, S. A. & Kasim, R. M. (2008a). Applications of multilevel models for meta-analysis. In *Multilevel Analysis of Educational Data*, A. A. O'Connell & D. B. McCoach (Eds.). Greenwich, CT: Information Age Publishing.

Kalaian, S., & Kasim, R. (2008b, October). *Why multivariate meta-analysis methods for studies with multivariate outcomes?* Paper presented at the annual meeting of the MWERA Annual Meeting, Westin Great Southern Hotel, Columbus, Ohio.

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

Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. (2009). The case of the misleading funnel plot. *BMJ, 333,* 597–600.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms*. (NCSER 2013-3000).

Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Lipsey, M., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist, 48,* 1181-1209.

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

Marsh HW, Bornmann L, Mutz R, Daniel H-D, and O'Mara A (2009). Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research, 79*(3), 1290-1326.

Marsh HW, & Bornmann L (2009). Do women have less success in peer review? *Nature, 459,* 602.

Marsh, H, & O'Mara, A (2008). Self-concept is as multidisciplinary as it is multidimensional: A review of theory, measurement, and practice in self-concept research. In H Marsh, R Craven, & D McInerney (Eds.) *Self-processes, learning and enabling human potential. International advances in self research.* Information Age, Charlotte, NC, pp. 87-115.

Mills E J, Thorlund K, Ioannidis JPA. (2013). Demystifying trial networks and network meta-analysis. *BMJ, 346,* f2914.

O'Mara, A. J. (2009). *Methodological and substantive applications of meta-analysis: Multilevel modelling, simulation, and the construct validation of self-concept* (Doctoral dissertation). University of Oxford, UK.

O'Mara, A. J., Marsh, H. W., Craven, R. G., & Debus, R. (2006). Do self-concept interventions make a difference? A synergistic blend of construct validation and meta-analysis. *Educational Psychologist, 41*, 181-206.

O'Mara-Eves, A. Brunton, G., McDaid, D., Oliver, S., Kavanagh, J., Jamal, F., Matosevic, T., Harden, A. and Thomas, J., (2013) Community engagement to reduce inequalities in health: a systematic review, meta-analysis and economic analysis. *Public Health Research*, 1 (4). pp. 1-548. ISSN 2050-4381

O'Muircheartaigh C, Hedges LV (2014) Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society: Series C 63*: 195-210.

Petticrew, M. (2015). Time to rethink the systematic review catechism? Moving from 'what works' to 'what happens'. *Systematic reviews, 4*(36), doi:10.1186/s13643-015-0027-1.

Ragin C. (2008a). *Redesigning social inquiry: Fuzzy sets and beyond.* London: University of Chicago Press.

Ragin C. (2008b). *User's guide to Fuzzy-Set / Qualitative Comparative Analysis*. Tucson, AZ: University of Arizona. Retrieved from http://www.u.arizona.edu/~cragin/fsQCA/download/fsQCAManual.pdf on 17 August 2015.

Raudenbush, S. W. (1994). Random effects model. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-321). New York: Russell Sage Foundation.

Raudenbush, S. W. & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10,* 75-98.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage

Riley, R, Lambert, PC, & Abo-Zaid, G (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ, 340,* c221

Riley, R. D., Thompson, J. R., & Abrams, K. R. (2008). An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics, 9*, 172-186.

Rosenthal R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638–41.

Sandelowski, M, Voils, CI, Barroso, J. (2006). Defining and designing mixed research synthesis studies. *Research in the Schools, 13*(1), 29-44.

Stanley, TD (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics, 70*, 103-127.

Sutton A (2009). Publication bias. In H Cooper, LV Hedges, & JC Valentine (eds.), *The handbook of research synthesis and meta-analysis (2nd edition),* New York: Russell Sage Foundation, pp. 435-452.

Sutton AJ, Song F, Gilbody SM, & Abrams KR. (2000). Modelling publication bias in meta-analysis: A review. *Statistical Methods in Medical Research, 9*, 421–445

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th Ed.). Pearson.

Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., et al. (2004). Integrating qualitative research with trials in systematic reviews: An example from public health. *BMJ, 328,* 1010–1012.

Thomas, J., O'Mara-Eves, A., & Brunton, G. (2014). Using Qualitative Comparative Analysis (QCA) in systematic reviews of complex interventions: A worked example. *Systematic Reviews, 3*, 67-81.

Thompson SG, Higgins JPT. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*, 1559-1574.

Thorlund K, Mills EJ. (2012). Sample size and power considerations in network meta-analysis. *Systematic Reviews, 1,* 41

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis.* Beverly Hills, CA: Sage.

Figure 1. Number of research publications with 'meta-analysis' in the title in the Web of Knowledge database, filtered by Education research area, over the last 40 years