Systematic review and meta-analysis

James Boyle, Michael Connolly & Tommy MacKay

Aim: *This paper provides an overview of the research methodologies of systematic review and meta-analysis and uses a commentary on the analysis of data from a previously published study to illustrate the procedures and decision-making involved for consumers and those who may be considering carrying out a systematic review.*

Rationale: *Systematic review and meta-analysis are located within a hierarchy of evidence-based practice, and their underlying epistemological and theoretical basis considered. The advantages of systematic review over traditional narrative reviews are discussed, together with the case for the use of meta-analysis to synthesise research findings. The feasibility of the use of these methodologies by educational psychologists is also considered.*

Findings: *The worked example details the steps necessary to carry out a systematic review and meta-analysis and the commentary addresses key issues such as specifying inclusion/exclusion criteria and determining relevance; specifying the literature search strategy and coping with the 'grey' literature; extracting and coding data from the included studies and the importance of reliability checks; study quality; selecting the most appropriate effect size; selecting the most appropriate model for meta-analysis (fixed-effect versus random-effect), combining and averaging effect sizes across studies; running weighted ANOVAs or meta-regression analyses to investigate heterogeneity; checks for publication bias; and sensitivity analysis to deal with outliers.*

Conclusions: *Future developments in these methodologies, details of available software and resources, and implications for educational psychologists who may wish to carry out systematic reviews and meta-analysis are discussed.*

Keywords: *Systematic review, meta-analysis, evidence-based practice.*

Systematic reviews and meta-analysis are examples of secondary data analysis (Glass (1976), that is, the re-analysis of existing original data from primary research studies carried out by other researchers. The approaches are linked to evidence-based practice and are widely used in psychology, education and health-related research to inform policy and professional practice, with the highest 'levels of evidence' commonly attributed to well-conducted systematic reviews and meta-analyses, as shown in Table 1 from the Scottish Intercollegiate Guidelines Network (2015).

[TABLE 1 ABOUT HERE]

Torgerson and Torgerson (2008) note that 'the main aim of a systematic review is to systematically locate either all of the available evidence on a given subject or a representative sample of the evidence, which may then be combined in a synthesis, such as a meta-analysis, in order to give a precise overview of the existing literature within an area'. Systematic reviews have grown in importance since the 1990s with the formation of The Cochrane Collaboration (http://uk.cochrane.org/) to promote and disseminate systematic reviews in health care, The Campbell Collaboration (http://www.campbellcollaboration.org) in 2000 with a similar remit in regard to research in education, social science and criminal justice, and the EPPI-Centre at the Institute of Education, University College London (http://eppi.ioe.ac.uk/cms/) established in 1993 with a remit for the synthesis of research in education, health, social care and society.

Systematic reviews are a response to concerns about the problems of traditional, descriptive narrative literature reviews. They provide clear research questions that they are designed to investigate, and explicit details of the search strategies and procedures used by

the reviewers to identify the literature (both published and unpublished). These include the criteria used to include (and exclude) studies from the review; the procedures of data extraction and coding; checks on external validity; and the techniques used to synthesise the data (Centre for Reviews and Dissemination, 2008). This approach has four advantages (Torgerson & Torgerson, 2008):

- providing the details of the methodology means that the review should be both transparent and replicable;
- the process of systematically identifying as complete a data set of independent studies as possible, published as well as unpublished, statistically significant as well as non-significant results, positive outcomes as well as negative or indeterminate, minimises sources of bias;
- providing the reasons for including and excluding studies prior to coding further minimises sources of bias;
- there is a methodology for synthesising the evidence across studies which provides a basis for testing the null hypothesis regarding the effectiveness of interventions.

Meta-analysis has a longer history than systematic reviews in general, with early developments driven by academic educational psychologists such as Gene Glass in the 1970s (Glass, 1976, 2000) who researched the effects of class size and also the effectiveness of psychotherapy. Statistical techniques of meta-analysis have become more refined over the last 40 years (Borenstein et al., 2009) and include analysis of variance and multiple regression to investigate the effects of moderator variables, such as which components of service delivery of implementation of an intervention are successful, or whether an intervention is more effective if targeted on a sub-group of participants. However, meta-analysis is not a requirement of a systematic review, and indeed, meta-analysis can be used to

synthesise data which has not been derived from a systematic review, for example, to inform the sample size for a primary research study (McCartney et al., 2004).

Implications for use by educational psychologists

Trainee educational psychologists undertaking doctoral studies in the UK receive training in carrying out a systematic review, with some universities providing weblinks to the reviews (for example, http://www.ucl.ac.uk/educationalpsychology/decpsy/evidencebased_practice_reports.html). In Scotland, trainees on the MSc in Educational Psychology at the University of Strathclyde commenced training in systematic review in 2015-16. This means that there is a steadily growing number of practitioner educational psychologists who have the necessary skills and expertise to carry out systematic reviews.

The three authors of this paper have recently completed two meta-analyses in a major study of autism commissioned by the Scottish Government. (An overview of the study, carried out in collaboration with researchers at the London School of Economics, is given in MacKay et al., (2013). The first study is a meta-analysis of the prevalence of autism spectrum disorders, while the second is a meta-analysis of distribution of intellectual ability/disability across the spectrum. Both studies are of crucial importance to planning future economic costings and development of provision for this sector which has become increasingly important in the work of educational psychologists. Prevalence and the presence or otherwise of intellectual disability are the two most significant factors in determining both the economic burden of autism at national level and the required future levels of supports and services (Knapp et al., 2009). A small over- or under-estimate of either of these variables can have vast budgetary and planning implications for education authorities and other agencies. Meta-

analyses, informed by the clinical insights of psychologists with expertise in this field, offer an important statistical tool in refining available estimates in studies of this kind.

However, while some research projects conducted by practitioners can be undertaken using a range of fairly basic statistical procedures which are widely accessible to most applied psychologists, the field of systematic review and meta-analysis is more complex in terms of its theoretical foundations and the statistical expertise required to apply its methods. The sections which follow cover theory and method, followed by a worked example of a relevant meta-analysis, and these sections should be particularly useful for those who are planning to carry out a study of this kind.

Theoretical/epistemological basis

Systematic review

The aim of a systematic review is to identify the best available evidence. However, as Hattie et al., (2014) point out, 'evidence is not neutral'. Table 1 reveals that traditional approaches to systematic reviews with their emphasis on randomised controlled trials (RCTs) or quasi-experimental designs and on generalisability of findings across different contexts are informed by positivist epistemology (Robson, 2011). However, critics (Biesta, 2007; Clegg, 2005) argue that such traditional approaches lead to a disjunction between 'scientific' understandings of 'effectiveness' and 'outcomes' on the one hand, and practice-based considerations informed by implementation science and understandings of the importance of context on the other (Dunst & Trivette, 2012).

Detailed protocols for carrying out systematic reviews of complex service interventions informed by critical realist epistemology (Robson, 2011) have been developed to provide frameworks for the synthesis of evidence using multiple methods and emphasising the importance of synthesising information about active programme ingredients and

implementation (American Psychological Association, 2008; Dunst & Trivette, 2012; Pawson et al., 2005), to answer the key questions 'What… works, for whom, in what circumstances, in what respects and why?' (Pawson et al., 2005).

These protocols draw important distinctions between 'intervention' and 'implementation' on the one hand, and 'practices' and 'outcomes' on the other (Dunst & Trivette, 2012). But informed by critical realism, the task of the systematic reviewer is to capture the details and nature of the practice considerations and conditions of implementation to identify underlying mechanisms which have associations with professional practices and outcomes. There are also interpretative approaches to the synthesis of qualitative data in the form of systematic narrative reviews of research based upon conceptual and thematic frameworks (Dunst & Trivette, 2012) for which detailed guidelines are available (Heaton, 1998; Popay et al., 2006; Snilstveit et al., 2012).

Meta-Analysis

The aim of meta-analysis is to synthesise quantitative data from studies included in the systematic reviews and to calculate overall aggregated or 'pooled' effect sizes (ES) and their associated confidence intervals. There are different types of ES which quantify the magnitude of treatment effects or between-group differences and in some cases provide information about the direction of the difference, as shown in Table 2 below. Here, we will focus on ES based upon means. Lipsey and Wilson (2001) and Borenstein et al. (2009, pp. 33-43) provide detailed discussion of ES based upon 2 x 2 tables and correlations.

[TABLE 2 ABOUT HERE]

ES based on means are widely used to compare outcomes from an intervention as they provide information not only about magnitude but also the direction of the effect (for example, whether it was positive). They can be calculated from pre- and post-intervention gain score data from one sample where there is no control group. They can also be calculated from comparisons between two groups, for example, comparing post-intervention means from intervention groups with those from control groups.

Standardised mean difference ES are commonly used because the division of the raw difference score by a standard deviation permits comparison between different outcome measures. The resulting metric, Cohen's *d* or Hedges' *g*, is also easily interpreted as it is a *z*-score. The basic formula for calculating Cohen's *d* for a comparison between two groups is:

$$ES = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}} = \frac{\overline{X}_1 - \overline{X}_2}{s_{pooled}}$$

Here, the mean post-intervention score for a control group ($\overline{X}_2$) is subtracted from the post-intervention score for an intervention group ($\overline{X}_1$). This raw-score difference is then divided by a pooled standard deviation, a weighted average of the standard deviations of the control group and the intervention group, which yields Cohen's *d* as a *z*-score. It is also possible to use the standard deviation from the control group rather than the pooled standard deviation to calculate *d* (see Lipsey and Wilson (2001) for a discussion).

Hedges' *g* is a variant of *d* which corrects for the bias occurring in the case of small sample size of less than 20. As an unbiased ES, it is routinely used in meta-analysis. Formulas also exist to derive *d* and *g* from t-values and degrees of freedom, F-values and degrees of freedom, and product moment correlations.

Meta-analysis provides a means of (a) pooling ES which capture magnitude and direction of effect across studies (with greater weight given to larger studies) to increase statistical power, and (b) investigating the relationships between outcomes and study variables of interest, which include those relevant to implementation of an intervention. However, it is important not to combine ES across meaningful categories. For example, ES derived from pre- versus post-intervention scores from one group are different in character from those derived from a comparison between two groups. Similarly, ES derived from small-N experimental designs are different in character from group designs. Care should therefore be taken not to combine ES across different study designs.

Empirical basis

Some examples of systematic reviews published by The Campbell Collaboration and The EPPI-Centre on a range of topics of interest to educational psychologists are shown in Table 3. None of these reviews was authored by a practitioner educational psychologist. Also shown in the table are details of three published reviews written by practitioner educational psychologists which were identified by a search of the PsycINFO database using the search term 'meta-analysis or systematic review' as methodology and 'educational psychology' as author affiliation. This yielded a sample of 118 papers of which two were authored by educational psychologists currently practising in the UK (Cole, 2008; Jones, 2013) together with a third more recent paper not yet on the database identified by hand-search (Randall & Tyldesley, 2016). Interestingly, all three papers were published in *Educational and Child Psychology*.

[TABLE 3 ABOUT HERE]

Example of systematic review and meta-analysis

We now turn to a commentary on a study carried out by the first author (Law et al., 1998). Following Uman (2011), we consider the key processes and decision-making involved in systematic review and meta-analysis in terms of a series of 'stages'.

1. Formulate the review question

The study was funded by the National Co-ordinating Council for Health Technology Assessment to investigate the literature regarding (a) the prevalence of primary speech and language delay (delays not attributed to hearing loss or other more general developmental disabilities) in the 0-7 years age-group (what is the scale of speech and language problems in young children); (b) its 'natural history' (what happens in the absence of intervention: do many children tend to 'catch up', with implications for the cost-effectiveness of intervention); (c) evidence for the effectiveness of intervention approaches (what can be shown to 'work', for whom, and in what circumstances); and (d) the accuracy of screening procedures (what proportion of children with language delay are incorrectly identified as 'false positives' and 'false negatives'). We focus here on the review questions for the intervention strand of the review:

- *What evidence is there that interventions can be shown to be effective when compared with untreated controls and other interventions?*

  The research literature showed that many young children identified with speech and language delay do indeed 'catch-up'. Accordingly, the reviewers elected to include in the review protocol only studies where there was control for the effects of maturation as a control for internal validity (that is, that observed change was associated with intervention).

- *For which sub-groups of children (characterised by age and communication skills) has intervention been shown to be most effective?*

  There was interest then as now in identifying whether particular interventions should be targeted on specific presenting problems.

- *What evidence is there for the role played by associated difficulties (for example, behaviour) in determining outcomes?*

  There was interest in the relationship between speech and language delay and behaviour difficulties and in whether behaviour was a moderator of outcomes for intervention.

- *Is there evidence that intervention for speech and language delay can be cost-effective?*

  There was interest in identifying and reviewing data relating to economic analyses of speech and language.

- *What components of the treatment process have an optimal effect?*

  There was interest in whether the 'active ingredients' of interventions could be identified.

- *Do effect modifiers mitigate against drawing useful comparisons between studies?*

  This review question reflected interest in whether the age of the child and the severity of presenting speech and language problems would make a difference to the outcome from intervention, for example, in regard to the case for early intervention rather than a 'wait and see' approach.

- *To what extent do the outcomes adopted reflect those recommended by the World Health Organisation (impairment, disability and handicap)?*

  There was interest within the speech and language therapy community about use of terminology.

2. Define inclusion and exclusion criteria

The inclusion criteria used in the study are given below, together with the rationale for their use:

- Reported after 1966 (that is, within the 30 year time period agreed with the sponsors: reviewers have to identify the time-period within which studies will be included) (*relevance*)
- Covers part of the age-range 0-7 years (*relevance)*
- Studies the effects of treatment/intervention upon speech and language delay in children (*relevance)*
- Study is of primary language delay (*relevance)*
- Details of the number of participants in each group are given *(statistical conclusion validity, external validity)*
- Provides a comparison of pre- and post-intervention speech and language measures *(internal validity, statistical conclusion validity)*
- Fulfils one of the following design criteria: (a) for group designs, experimental study with randomised non-treatment controls or quasi-experimental studies with non-equivalent non-treatment controls; (b) for single-subject experimental designs, withdrawal and reversal designs, multiple baseline designs, multiple probe designs or alternating treatment designs *(internal validity)*
- Provides details of the nature, duration, span and delivery of treatment (linking into implementation) *(construct validity, internal validity)*

3. Develop search strategy and locate studies

Published, unpublished and 'grey' (for example, reports, theses etc.) literature from the period January 1967 to May 1997 was searched for relevant studies. Searches were carried out on the six databases that Cros and DialIndex searches indicated were most relevant for

this subject area, namely CINAHL (Cumulative Index of Nursing and Allied Health), Embase, ERIC (Educational Resources International Clearing House), LLBA (Linguistics and Language Behaviour Abstracts), Medline and PsycLIT. Two databases dealing with unpublished literature (SIGLE and Boston Spa Conferences) were also checked. In addition, bibliographies from compilation volumes and articles retrieved and internet sources were checked, key journals were hand-searched and calls for information were made to professional organisations, institutions and authors. (Full details of the literature retrieval process may be found in Law et al., 1998).

4. Select studies

In all, 125 relevant intervention papers were identified from the period 1967-97. Of these, 45 papers met the above inclusion criteria. 19 papers provided 21 datasets from group designs: 10 RCTs and 11 from quasi-experimental designs. 26 papers provided 26 datasets from single-subject (small-n) experimental designs.

5. Extract data

Studies which met the inclusion criteria were coded using a data extraction form (see Law et al., 1998, pp. 154-156). Two independent assessors made final judgements about inclusion. Further details together with the reason for rejecting any relevant study which failed to meet the above criteria may be found in Law et al. (1998).

6. Assess study quality

All studies included in the review were rated for quality in terms of reliability (defined here as the information required for replication) and validity (internal and external). Factors considered to be of importance to validity in the case of group designs included

subject recruitment, subject allocation to groups, reporting of subjects' speech and language profile and blinding of assessors. Reliability factors noted included reporting of the subjects' details (for example, age, gender, socioeconomic status, ethnicity), treatment details and assessment instruments. Gough (2007) provides a useful framework for evaluating study quality and weight of evidence.

All coding of the intervention studies was carried out by the first author of the present paper. A sample of 4 of the 47 intervention studies which met the inclusion criteria (9%) was coded independently by two of the systematic review co-authors. The overall percentage agreement rate between the two independent coders was 89% (range 62.5%-100% per item). The most common source of lack of agreement was omission by coders of minor details of a kind that would have been resolved by discussion (for example, calculation of standard deviations using the formula for populations rather than that for samples). A further confirmatory point-by-point check on a second sample of 5 papers (11% of the total number of included studies) yielded 99% agreement.

The results from the group design intervention studies were synthesised by converting the outcomes into effect sizes to permit comparison across studies. In the case of the RCT and quasi-experimental group designs, the effect size used was Hedges' *g*, the difference between the post-test means of treatment and non-treatment control groups divided by the pooled standard deviation for each study, corrected for population effect size bias. Effect sizes with positive signs indicated that subjects in treatment groups achieved higher post-intervention scores that those in non-treatment control groups, that is, that there was a positive treatment effect. Conversely, effect sizes with negative signs indicated studies in which subjects in treatment groups failed to make greater progress than those in control groups.

7. Analyse and interpret results

Of the 21 included datasets with group designs, we present an analysis of the largest subset, those with outcomes for intervention in expressive language to illustrate the procedures involved in meta-analysis. There were 14 such datasets (see the references at the end of the paper) which were re-analysed for this paper using Comprehensive Meta-Analysis v. 3.3.070 (https://www.meta-analysis.com/).

The first decision is whether to use a fixed- or random-effect model for the meta-analysis. Our original analysis was based on a fixed-effect model, common practice amongst researchers at the time. A fixed-effect model assumes that ES are homogeneous, with one common or 'true' effect size in all studies and only one source of error, random sampling error from a single population of studies (i.e. weighted within-studies variance) (Borenstein et al., 2009). In contrast, the more conservative random-effect model assumes that individual ES are heterogeneous because they are randomly sampled from a 'universe' of populations (Hattie et al., 2014). This means that there are two sources of sampling error in a random-effect model: weighted within-study variance plus an estimate of the population variance. We selected a random-effect model for this re-analysis.

The aim of a meta-analysis of intervention studies is to combine and average effect sizes across studies and to use confidence intervals and homogeneity statistics to calculate whether the average effect size is significantly different from zero. As a precursor, it is good practice to examine a funnel plot. Funnel plots are used to determine publication bias, that is, whether there are significant numbers of 'missing' studies which have not been published or otherwise reported as a result of non-significant findings. Non-significance is an important consideration for studies which investigate intervention effects as publication bias might inflate estimates of the effects of an intervention. But funnel plots also provide information about statistical outliers which contribute to the

heterogeneity to be explained or accounted for in the meta-analysis. A funnel plot consists of a graph with the ES for each study in the X axis plotted against its standard error with the 95% confidence intervals (CI) as shown in Figure 1. If there is no publication bias, then studies should be distributed around the mean effect. However, if there is publication bias, then there should be gaps in the left of the funnel plot where non-significant findings would be located (Borenstein et al., 2009). Estimates outwith the 95% CI may also indicate possible statistical outliers.

[FIGURE 1 ABOUT HERE]

The funnel plot for the 14 ES reveals no indication of publication bias (Begg and Mazumdar Rank Correlation Test Kendall's tau b = 0.242, 1-tailed p-value = 0.114; Egger's Intercept = 0.202, 95% CI (-1.430, 1.834), 1-tailed p-value = 0.396), although the small number of studies should be noted. A larger number of studies would have increased the statistical power of these tests. However, there are two possible outliers: the +2.38 ES from the Wilcox and Leonard (1978) study and the +1.85 ES from Gibbard (1994) first study.

A non-parametric DerSimonian-Laird random effects model revealed a large statistically significant overall pooled ES of +0.913 (95% CI 0.643, 1.184), Z = 6.62, p = .000, 2-tailed). However, significant heterogeneity was indicated by a weighted within-groups analysis of variance, the Q-statistic, which was statistically significant (Q = 24.19, d.f. = 13, p = .029, 2-tailed). This was confirmed by the I^2 statistic, which indicates moderate heterogeneity, and specifically that 46.26% of the variance results from differences in ES that could be explained by candidate moderator analysis.

We first considered the continuous moderator variables in the data set. These were child's age, reliability score for the study as rated by the reviewers, validity score for the study as

rated by the reviewers, gender balance of the sample, number of hours of intervention, number of sessions of intervention and number of months of intervention. With the relatively small number of studies and a recommended ratio of 10 studies for each variable (Borenstein et al*., 2009)*, each of these variables was entered separately into an analysis rather than in combination in a meta-regression.

The results revealed that none of these variables accounted for significant amounts of variance (all Q-values < 1.81, d.f. = 1, all p-values > .181, 2-tailed). Similar results were obtained from the categorical variables of study design (RCT versus quasi-experimental study design); intervention delivered by a trained therapist versus delivery by a parent or a teacher; norm-referenced versus criterion referenced measures of outcome and type of intervention ('naturalistic' versus 'didactic' versus 'hybrid'). Again, each variable was entered separately. Again, the results revealed that none of these variables accounted for significant amounts of variance (all Q-values < 0.711, d.f. = 1, all p-values > .400, 2-tailed).

A further random-effect meta-analysis was carried out with the Wilcox and Leonard (1978) outlier study removed from the analysis. This is called a 'sensitivity analysis' (Borenstein et al*.,* 2009) and is designed to investigate whether the remaining studies might be homogeneous in the absence of the outlier. The findings from this are shown in Table 4. The table displays Hedges' *g* ES (with its associated 95% CI) for each study, together with a z-value and associated p-value to indicate whether the intervention was successful in that study. The ES and 95% CI are also graphed for each individual study.

[TABLE 4 ABOUT HERE]

This is called a 'forest plot'. If a lower-bound 95% CI in a forest plot crosses zero here, then the intervention group did not achieve statistically significantly higher post-intervention

scores than the control group. In the case of this data set of 13 studies, only six reported statistically significant outcomes in favour of the intervention, which may be associated with the small sample sizes shown in Table 4. However, the strength of a meta-analysis is that it achieves greater statistical power by pooling across studies, as evidenced by the pooled ES of +0.838 (95% CI 0.617, 1.059), Z = 7.421, p = .000, 2-tailed) shown at the end of the table. The non-significant Q statistic of 15.30, d.f. = 12, p =. 225, 2-tailed, the associated tau^2 statistic of 0.033 and I^2 statistic of 21.58% all confirm that there is no further significant heterogeneity to be accounted for once the Wilcox and Leonard (1978) study has been removed. We can therefore treat the ES of +2.28 from the Wilcox and Leonard study as an outlier and conclude that the statistically-significant pooled Hedges' *g* ES of +0.838 (95% CI 0.617, 1.059) from a random-effect model is an appropriate overall estimate of the intervention effects for expressive language in this study, concluding further that intervention for expressive language delay in young children relative to control groups can be effective.

We also conclude that the following factors were not associated with the outcomes from the included studies here: (a) study design (whether the study was an RCT with random allocation of participant to intervention and control group or a quasi-experimental design with non-random allocation); (b) instrumentation (whether the outcome measure was a standardised test score or a criterion-referenced measure); (c) the delivery of intervention (for example, by a therapist or by a parent or teacher); (d) the specific type of intervention; (e) the age of the child or gender balance of the participants; (f) reliability and validity scores for the studies (all high); or (g) number of sessions/hours/duration of the programme. In terms of delivery and cost-effectiveness, the data further suggest that parents and teachers can be as effective agents of change in delivering programmes of intervention designed by speech and language therapists as the therapists themselves. However, the small number of studies and an absence of data for some of the variables relating to implementation should be noted.

Some additional points to note are that no more than one ES per participant in a sample can be included in a meta-analysis. If there are multiple measures, then there are two ways forward. The first is to collapse across the multiple measures and calculate an average ES for each participant and include that in the analysis. This assumes that it is meaning to combine the specific scores in question and that they should be measuring the same construct. The second approach is to carry out an 'upper-bound' analysis of the highest ES from each study followed by a 'lower-bound' analysis of the lowest ES from each study and to compare the results to see if there is a difference.

Finally, there are also Bayesian approaches to meta-analysis which are beyond the score of this overview. These yield wider confidence intervals (referred to as 'credibility limits') than even random-effect models and raise concerns that prior probabilities may be based on subjective assumptions (see Sutton and Abrams (2001) for a discussion).

Conclusion

While systematic reviews and meta-analysis can make important contributions to evidence-based practice and to our understanding of the relationships between underlying factors, there are limitations which should be noted. While systematic reviews are regarded by some researchers, particularly those in health-related fields, as the 'gold-standard' for efficacy and effectiveness research and for informing policy, they have traditionally privileged the use of RCT and other experimental designs. However, these designs emphasise both de-contextualised approaches to intervention and high levels of researcher control, which may run counter to understandings of the importance of empowerment and emancipation in achieving change from complex interventions in open systems (Clegg, 2005; Snilstveit et al., 2012).

Further, only the available data can be analysed. This means that there may be small numbers of studies relevant to given research questions. And as we have seen, there may also be an absence of data for key variables relating to implementation, which constrains moderator analysis. Locating all available and relevant data is also a limitation and is linked to the problem of publication bias. Studies showing effective interventions are more likely to be published and hence accessible than unpublished studies with non-significant effects, which may lead to inflated, overestimated ES, which might in turn lead to over-confidence in the effectiveness of given interventions.

But systematic reviews of the literature have come to the fore in recent years by adding a degree of replicability which is difficult to achieve with traditional, narrative reviews. Allied to transparent procedures for locating and evaluating available and relevant evidence, this can make a significant contribution to minimising sources of bias.

The parallel development of the techniques of meta-analysis permits the pooling of effects across studies identified by systematic reviews to enhance the statistical power of the overall analysis. As Table 4 revealed, only six of the 13 studies included in the final meta-analysis model achieved conventional levels of significance when considered individually. But aggregating the weighted ES across studies generated a pooled sample size of 323 (165 participants from intervention groups and 158 from control groups), which markedly increases the power of the analyses. The statistical tools for meta-analysis in turn allow the analyst to investigate sources of heterogeneity in pooled ES by analysis of moderators by ANOVA and multiple regression. This can help to illuminate issues relevant to the implementation of interventions, although it is possible to draw inferences only about associations, not about causality.

Time constraints and access to the professional and academic literature for those without a university affiliation have been major barriers for educational psychologists wishing to carry

out systematic reviews. But while the time demands should not be minimised, the recently introduced requirement in UK universities of the Research Excellence Framework (REF) that UK academics must publish in open access journals is a positive development. In the medium term at least, practitioners will have access to increasing numbers of papers which could form the basis for systematic reviews.

Systematic reviews and meta-analysis are not without their critics, but they may be regarded as tools which help to make sense of complex information. It is also likely that they may become more relevant to practitioner educational psychologists, as the methodologies become informed by the use of protocols closely linked to implementation science and informed by a critical realist perspective and are linked in turn to understandings derived from implementation science (American Psychological Association , 2008; Dunst and Trivette, 2012; Pawson et al., 2005).

Future directions

In terms of future directions, there continue to be developments to approaches which make 'artifact' adjustments to effect sizes where there are problems with reliability of measurements or range restriction (Schmidt & Hunter, 2015). These adjustments are designed to provide additional means of correcting error and bias, and while some commentators argue that they yield 'idealised' rather than observed effect sizes, the approaches may be worth considering.

Systematic reviews and meta-analysis are well-served by resources. The coding templates developed by a task force of educational psychologists in the US (American Psychological Association, 2008) for use with group designs and single-case experimental small-N designs may prove helpful to those interested in evaluating the effectiveness of the delivery of complex programmes of intervention. The PRISMA Group (Moher, Liberati, Tetzlaff,

Altman, & The PRISMA Group, 2009) also provide a useful template for a flow diagram to document a systematic review.

The Cochrane Collaboration produce 'RevMan 5.3' (http://tech.cochrane.org/revman), a programme to support systematic reviews and meta-analysis and to maintain and update existing reviews. The software is free with on-line tutorials and the PRISMA flow chart is completed as part of the process. However, support is only available to those registered with the Cochrane Collaboration. In a similar vein, the EPPI-Centre website (http://eppi.ioe.ac.uk/cms/) provides details of a useful mapping strategy (used to check that studies are suitable for inclusion), a keywording strategy, data extraction guidelines and the Review Group Manual. The Centre has also developed procedures for triangulating sources of qualitative data and for synthesising outcomes based upon qualitative data.

In addition to RevMan 5.3, Stata (http://www.stata.com/), Comprehensive Meta Analysis (https://www.meta-analysis.com/) and MetaWin (http://www.metawinsoft.com/) are fully-featured stand-alone programmes designed to carry out all aspects of a meta-analysis. MetaEasy (http://www.statanalysis.co.uk/meta-analysis.html) and MetaXL v5.1 (http://www.epigear.com/index_files/metaxl.html) are add-in programmes for Excel which are worth exploring. MetaFor (http://www.metafor-project.org/doku.php) is open-source software for the programme language R and there are also add-on scripts for SPSS available from the following websites: http://mason.gmu.edu/~dwilsonb/ma.html and http://www.statisticshell.com/meta_analysis/how_to_do_a_meta_analysis.html.

Finally a recent development which may be of interest to those planning large-scale systematic reviews with the potential for publication is PROSPERO, an international database for protocols which can be peer-reviewed and registered before the review commences (http://www.crd.york.ac.uk/PROSPERO/).

Suggested readings

Borenstein, M., Hedges, L.V., Higgins, J.P.T. & Rothestein, H.R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley & Son.

Cooper, H., Hedges, L.V. & Valentine, J.C. (2009). *The handbook of research synthesis and meta-analysis* (2nd Edn). New York: Russell Sage Foundation.

Field, A.P. (2003). Can meta-analysis be trusted? *The Psychologist,* 16(12), 642-645.

Field, A.P. & Gillett, R. (2010). How to do a meta-analysis. British Journal of Mathematical and Statistical Psychology, 63(3), 665-694.

Lipsey, M.W. & Wilson, D.B. (2001). *Practical meta-analysis*: London: Sage.

Acknowledgements

The systematic review and meta-analysis reported here were funded by the National Co-ordinating Council for Health Technology Assessment (Project No. 99/36/04) to James Law and the first author. The views expressed in this paper are those of the authors alone and not necessarily those of the HTA programme or the UK Department of Health.

Address for correspondence:

Professor James Boyle

School of Psychological Sciences and Health,

University of Strathclyde,

40 George Street,

Glasgow G1 1QE.

Email: j.boyle@strath.ac.uk

References to studies included in the meta-analysis example

Almost, D. & Rosenbaum, P. (1998). Effectiveness of speech intervention for phonological disorders: a randomized controlled trial. *Developmental Medicine and Child Neurology*, *40*, 319–52.

Fey, M.E., Cleave, P.L., Long, S.H. & Hughes, D.L. (1993). Two approaches to the facilitation of grammar in children with language impairment: an experimental evaluation. *Journal of Speech and Hearing Research*, *36*, 141–57.

Fey, M.E., Cleave, P.L., Ravida, A.I., Long, S.H., Dejmal, A.E. & Easton, D.L. (1994). Effects of grammar facilitation on the phonological performance of children with speech and language impairments. *Journal of Speech and Hearing Research*, *37*,594–607.

Gibbard, D. (1994). Parental-based intervention with pre-school language-delayed children. *European Journal of Disorders in Communication*, *29*(2), 131–150.

Girolametto, L., Pearce, P.S. & Weitzman, E. (1996). Interactive focused stimulation for toddlers with expressive vocabulary delays. *Journal of Speech and Hearing Research*, *39*(6), 1274–83.

Girolametto, L., Pearce, P.S. & Weitzman, E. (1995). The effects of focused stimulation for promoting vocabulary in young children with delays: a pilot study. *Journal of Child Communication Development*, *17*(2), 39–49.

McDade, A. & McCartan, P.A. (1996). *Partnership with parents: A pilot study.* Edinburgh: Monklands and Cumbernauld Division of Speech and Language Therapy, Report to the Scottish Office Home and Health Department.

Matheny, N. & Panagos, J.M. (1978). Comparing the effects of articulation and syntax programs on syntax and articulation improvement. *Language, Speech and Hearing Services in Schools*, *9*, 57–61.

Schwartz, R.G., Chapman, K., Terrell, B.Y., Prelock, P. & Rowan, L. (1985). Facilitating word combination in language-impaired children through discourse structure. *Journal of Speech and Hearing Disord*ers, *50*, 31–39.

Stevenson, P., Bax, M. & Stevenson, J. (1982). The evaluation of home-based speech therapy for language delayed preschool children in an inner city area. *British Journal of Disorders in Communication*, *17*(3), 141–148.

Whitehurst, G.J., Fischel, J.E., Lonigan, C.J, Valdez-Menchaca, M.C., Arnold, D.S. & Smith, M. (1991). Treatment or early expressive language delay: if, when, and how. *Topics in Language Disorders*, *11*(4), 55–68.

Wilcox, M.J. & Leonard L.B. (1978). Experimental acquisition of Wh- questions in language-disordered children. *Journal of Speech and Hearing Research*, *21*, 220–239.

Zwitman, D.H. & Sonderman, J.C. (1979). A syntax program designed to present base linguistic structures to language-disordered children. *Journal of Communication Disorders*, *12*(4), 323–35.

References

American Psychological Association. (2008). *Procedural and coding manual for review of evidence-based interventions* (2nd ed.). Retrieved from http://www.indiana.edu/~ebi/projects.html

Biesta, G. (2007). Why 'What Works' Won't Work: Evidence-Based Practice and the Democratic Deficit in Educational Research. *Educational Theory, 57*(1), 1-22.

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

Centre for Reviews and Dissemination. (2008). *Systematic reviews: CRD's guidance for undertaking reviews in health care* (3rd ed.). University of York: Author.

Clegg, S. (2005). Evidence-based practice in educational research: a critical realist critique of systematic review. *British Journal of Sociology of Education, 26*(3), 415-428.

Cole, R.L. (2008). A systematic review of cognitive-behavioural interventions for adolescents with anger-related difficulties. *Educational and Child Psychology, 25*(1), 27-47.

Dunst, C.J., & Trivette, C.M. (2012). Meta-Analysis of Implementation Practice Research. In B. Kelly and D. Perkins (eds.) *Handbook of implementation Science* (pp. 68-91). Cambridge: Cambridge University Press.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8.

Glass, G.V. (2000). Meta-analysis at 25. Retrieved from www.gvglass.info/papers/meta25.html

Gough, D. (2007). Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education, 22*(3), 213-228.

Hattie, J., Rogers, H.J. & Swaminathan, H. (2014). The role of meta-analysis in educational research. In A. D. Reid, P. E. Hart & M. A. Peters (eds.) *A Companion to research in education* (pp. 197-207). Dordrecht, Netherlands: Springer Science+Business Media.

Heaton, J. (1998). Secondary analysis of qualitative data. *Social Research Update 22.* Retrieved from http://sru.soc.surrey.ac.uk/SRU22.html

Jones, T.W. (2013). Equally cursed and blessed: Do gifted and talented children experience poorer mental health and psychological well-being? *Educational and Child Psychology, 30*(2), 44-66.

Knapp, M., Romeo, R. & Beecham, J. (2009). Economic cost of autism in the UK. *Autism, 13*(3), 317–336.

Law, J., Boyle, J., Harris, F., Harkness, A. & Nye, C. (1998). Screening for speech and language delay: a systematic review of the literature. *Health Technology Assessment, 2*(9), 1-184.

Lipsey, M.W. & Wilson, D. B. (2001). *Practical meta-analysis*. London: Sage.

MacKay, T., Boyle, J., Knapp, M. & Connolly, M. (2013). A multi-strand investigation of microsegmentation of the autism spectrum to enhance the data on the economic costs and benefits of provision. *Good Autism Practice, 14,* Supplement 1, 101-106.

McCartney, E., Boyle, J., Bannatyne, S., Jessiman, E., Campbell, C., Kelsey, C., Smith, J. & O'Hare, A. (2004). Becoming a manual occupation? The construction of a therapy manual for use with language impaired children in mainstream primary schools. *International Journal of Language and Communication Disorders, 39*(1), 135-148.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Medicine 6*(7)*: e1000097.*

Pawson, R., Greenhalgh, T., Harvey, G. & Walshe, K. (2005). Realist review--a new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy, 10 Suppl 1*, 21-34.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N. with Roden, K. and Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC methods programme. Retrieved from http://www.lancaster.ac.uk/shm/research/nssr/research/dissemination/publications/NS_Synthesis_Guidance_v1.pdf

Randall, L. & Tyldesley, K. (2016). Evaluating the impact of working memory training programmes on children – a systematic review. *Educational and Child Psychology, 33*(1), 34-50.

Robson, C. (2011). *Real world research. 3rd. edn.* Oxford: John Wiley & Sons.

Schmidt, F.L. & Hunter, J.E. (2015). *Methods of meta-Analysis: correcting error and bias in research findings*. London: Sage Publications Ltd.

Scottish Intercollegiate Guidelines Network. (2015). SIGN 50: A guideline developer's handbook. Quick reference guide. *Author* (pp. 1-19). Edinburgh.

Snilstveit, B., Oliver, S. & Vojtkova, M. (2012). Narrative approaches to systematic review and synthesis of evidence for international development policy and practice. *Journal of Development Effectiveness, 4*(3), 409-429.

Sutton, A.J. & Abrams, K.R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research, 10*(4), 277-303.

Torgerson, D.J. & Torgerson, C.J. (2008). *Designing randomised trials in health, education and the social sciences: An introduction.* Basingstoke, Hampshire: Palgrave MacMillan.

Uman, L.S. (2011). Systematic Reviews and Meta-Analyses. *Journal of the Canadian Academy of Child and Adolescent Psychiatry, 20*(1), 57-59.

Table 1: 'Levels of Evidence' (Scottish Intercollegiate Guidelines Network, 2015)

Levels of Evidence	
1++	High quality meta-analyses, systematic reviews of randomised controlled trials (RCTs) or RCTs with a very low risk of bias
1+	Well conducted meta-analyses, systematic reviews or RCTs with a low risk of bias
1 -	Meta-analyses, systematic reviews or RCTs with a high risk of bias
2++	High quality systematic reviews of case control or cohort studies High quality case control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal
2+	Well conducted case control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal
2 -	Case control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal
3	Non-analytic studies, e.g. case reports, case series
4	Expert opinion

Table 2: Examples of Different Types of Effect Size (adapted from Borenstein et al. **(2009)**

ES based on means	**ES based on binary data (2 x 2 tables)**	**ES based on correlational data**
Raw (unstandardised) mean difference (ie gain score) Standardised mean difference (Cohen's *d* or Hedges' *g*)	Proportions (e.g. for analyses of prevalence) Risk ratio Odds ratio Risk difference	Pearson's product moment correlation

Table 3: Some examples of published systematic reviews relevant to the work of educational psychologists

The Campbell Collaboration Library of Systematic Reviews (http://www.campbellcollaboration.org/lib/)
Farrington, D.P. & Ttofi, M.M. (2009). School-based programs to reduce bullying and victimization. *Campbell Systematic Reviews 2009:*6.
Morton, M. & Montgomery, P. (2011). Youth empowerment programs for improving self-efficacy and self-esteem of adolescents. *Campbell Systematic Reviews 2011:*5.
Nye, C., Schwartz, J. & Turner, H. (2006). Approaches to parent involvement for improving the academic performance of elementary school age children. *Campbell Systematic Reviews 2006:*4.
Oliver, R., Wehby, J. & Daniel, J. (2011). Teacher classroom management practices: Effects on disruptive or aggressive student behaviour. *Campbell Systematic Reviews 2001:*4.
Wilson, S.J., Tanner-Smith E., Lipsey, M., Steinka-Fry, K.T. & Morrison, J. (2011). Dropout Prevention and Intervention Programs: Effects on school completion and dropout among school-aged children and youth. *Campbell Systematic Reviews 2011:*8.
Zief, S.G., Lauver, S. & Maynard, R.A. (2006). Impacts of After-school programs on student outcomes. *Campbell Systematic Reviews 2006:*3.
The EPPI-Centre (http://eppi.ioe.ac.uk/)
Cajkler, W., Tennant, G., Tiknaz, Y., Sage, R., Tucker, S. & Taylor, C. (2007). A systematic literature review on how training and professional development activities impact on teaching assistants' classroom practice (1988-2006). In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Garcia, J., Sinclair, J., Dickson, K., Thomas, J., Brunton, J., Tidd, M. & the PSHE Review Group (2006). Conflict resolution, peer mediation and young people's relationships. Technical report. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Kavanagh, J., Oliver, S., Caird, J., Tucker, H., Greaves, A., Harden, A., Oakley, A., Lorenc, T. & Thomas, J. (2009). *Inequalities and the mental health of young people: A systematic review of secondary school-based cognitive behavioural interventions.* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Nind, M., Wearmouth, J., with Collins, J., Hall, K., Rix, J. & Sheehy, K. (2004). A systematic review of pedagogical approaches that can effectively include children with special educational needs in mainstream classrooms with a particular focus on peer group interactive approaches. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Sheehy, K. & Rix, J. (2009). A systematic review of whole class, subject-based pedagogies with reported outcomes for the academic and social inclusion of pupils with special educational needs. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Thomas, J., Vigurs, C., Oliver, K., Suarez, B., Newman, M., Dickson, K. & Sinclair, J. (2008). Targeted youth support: Rapid evidence assessment of effective early interventions for youth at risk of future poor outcomes. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Some examples of recently published systematic reviews by educational psychologists
Cole, R.L. (2008). A systematic review of cognitive-behavioural interventions for adolescents with anger-related difficulties. Educational and Child Psychology, *25*(1)*,* 27-47.
Jones, T.W. (2013). Equally cursed and blessed: Do gifted and talented children experience poorer mental health and psychological well-being? Educational and Child Psychology, *30*(2)*,* 44-66.
Randall, L. & Tyldesley, K. (2016). Evaluating the impact of working memory training programmes on children: A systematic review. Educational and Child Psychology, *33*(1), 34-50.

Figure 1: Funnel plot of standard error by Hedges' *g* expressive language outcomes in the 27-73 months age-range from group design studies (N=14)

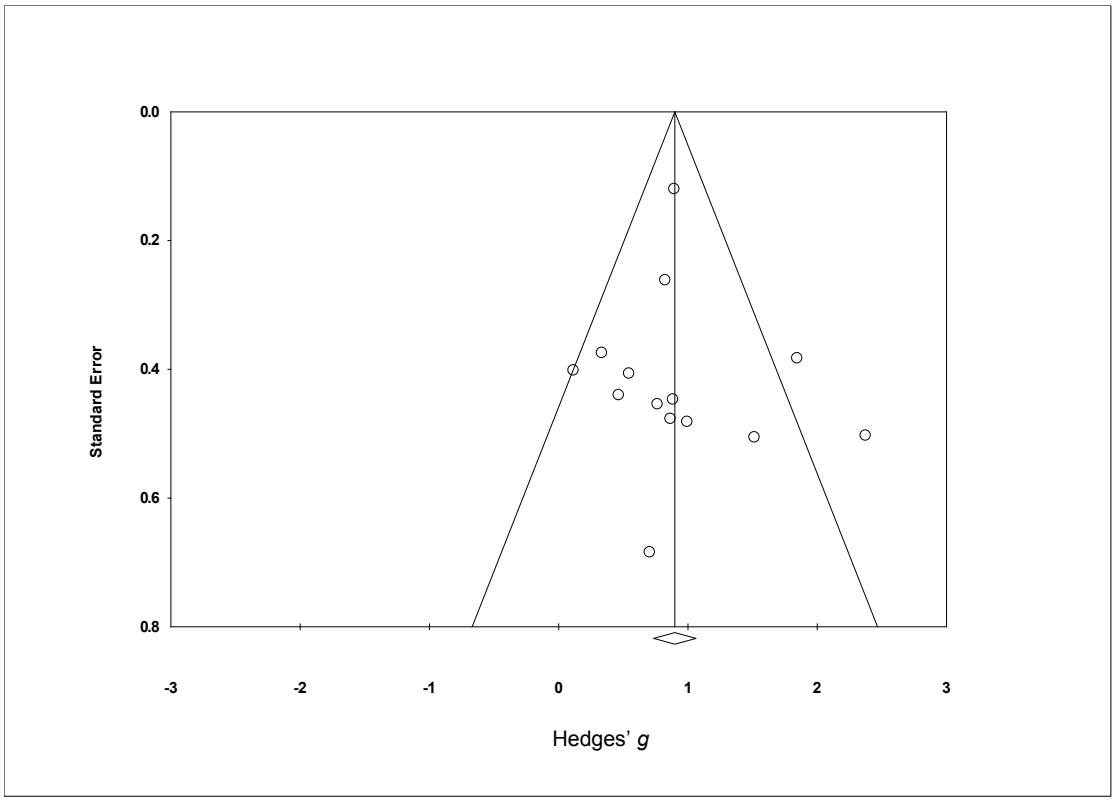
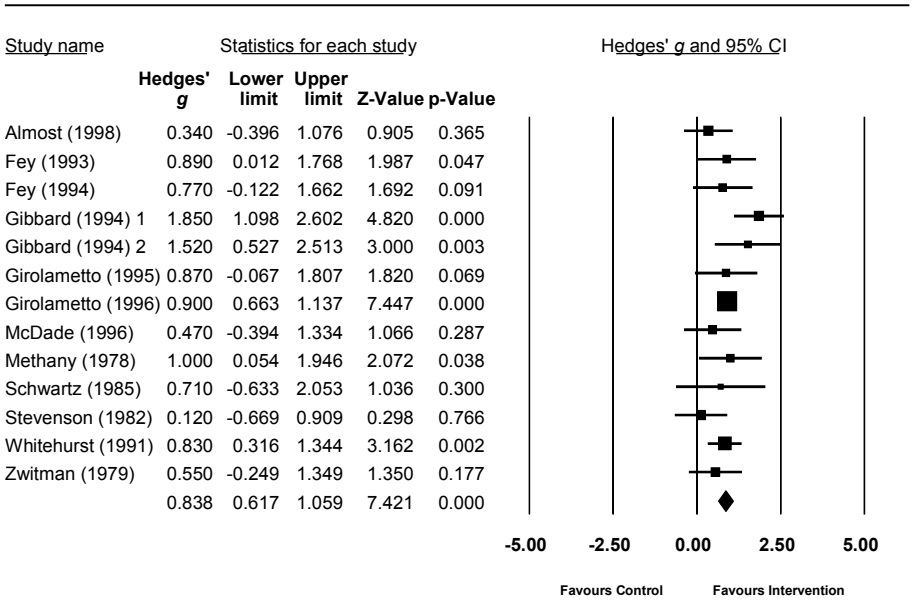


Table 4: Findings from a meta-analysis of expressive language outcomes in the 27-73 months age-range following the intervention



Q = 15.302, d.f. = 12, p = .225, I-Sq = 21.583%, Tau-Sq = 0.033