30

SYSTEMATIC REVIEWS

Carole Torgerson, Jill Hall and Kate Light

Overview

Systematic reviews are rigorously designed and conducted literature reviews that aim to exhaustively search for, identify, appraise the quality of and synthesise all the high-quality research evidence in order to answer a specific research question. Systematic reviews are designed to limit all potential sources of bias in reviewing a body of literature.

Introduction

Traditional literature reviews

Literature reviews seek to consolidate existing theoretical and empirical knowledge on specific issues. 'Traditional' literature reviews, sometimes termed 'narrative' or 'expert' reviews, are generally based on expert substantive knowledge in a given area. Generally, there is little or no clear rationale for the design and methods of such reviews. Typically, an expert in a substantive topic area gathers together and interprets previous research in the field and draws conclusions about the studies selected. However, the selection of studies for inclusion is usually not explicit, and whether the included studies are a truly representative or a 'biased' sample of the existing literature is often not clear. There are a number of potential problems with traditional literature reviews, including pre-existing author bias towards a particular hypothesis, which may in turn lead to a biased review.

Systematic reviews

A systematic review has been defined as '. .. the application of strategies that limit bias in the assembly, critical appraisal and synthesis of all relevant studies on a given topic' (Chalmers et al., 2002). The philosophy underpinning systematic review design is based on the scientific principle of replication. Systematic reviews are designed to be explicit, transparent and replicable in order to overcome many of the potential problems associated with the design of traditional reviews. If a review is to be replicable it needs to be explicit about how the various studies included in the review were identified and synthesised. All assumptions and reviewer judgements are made explicit and open to scrutiny and replication. Systematic reviews also seek to search exhaustively for all the relevant studies, whether formally published or listed in the 'grey' literature, and to include the 'totality' of studies in a field. Therefore systematic review design is less likely to suffer from reviewer selection bias. In addition, the exhaustive nature of the review process offers some protection against other forms of potential bias, in particular publication bias (see below).

Systematic reviews have a long history, with some of the first being reported in astronomy more than 100 years ago (Petticrew, 2001; Chalmers et al., 2002). Glass first invented meta-analysis, a statistical method for combining similar studies, for use in the field of education/psychology in the 1970s (Glass, 1976; Glass et al., 1981), and he pioneered the use of systematic reviews and meta-analysis in the field of education. After a period in which systematic reviews and meta-analyses tended to fall out of use, in the last 20 years or so their use has increased in prominence, first in the field of healthcare research and more recently in education and the social sciences.

Focus of this chapter

Systematic review methodology can be used to inform the design of a number of types of review. Scoping reviews can map out the research in a field while tertiary reviews can locate, critically appraise and synthesise existing systematic reviews in a field. Systematic reviews vary in emphasis in terms of their design and the inclusion of studies selected for specific kinds of research questions. It should be noted that systematic reviews can answer questions of 'why?' or 'how?', where it might be appropriate to identify empirical research using qualitative designs. Much of the information on the design and methods of systematic reviews contained within this chapter can be applied to systematic reviews of this nature. However, this chapter focuses on effectiveness questions and therefore on experimental research, as those studies most likely to be included in systematic reviews address these types of questions. These designs offer the potential of a counterfactual to demonstrate what would have happened to the participants had the intervention not been introduced. Ideally the same school, class or group of individuals would be observed under one condition and then observed again under the alternative condition. However, this is generally not possible (except in the relatively unusual circumstances of a cross-over trial). Consequently it is necessary to assemble two or more groups, with one group receiving the intervention and the other receiving an alternative intervention or 'business as usual'. It is then possible to compare the groups to see if there are any differences and potentially ascribe these differences to the intervention under evaluation.

Systematic review design and methodology

The rationale for systematic reviews focuses on the key principles of objectivity and scientific rigour. Systematic review design enables potentially unmanageable amounts of literature to be managed in a scientifically credible and reliable way and it enables the consistency and generalisability of research findings to be tested and all potential sources of bias to be minimised (Mul-row, 1994; Chalmers et al., 2002).

Systematic reviews use explicit methods to locate, appraise the quality of and synthesise the results of relevant research. To minimise the risk of bias the methods are pre-defined. This is important because once studies are identified it is critical that the inclusion/exclusion criteria are not changed in order to support a hypothesis that has been developed through exposure to some of the studies identified. There is a consensus regarding the design, methodology and methods of systematic reviews, a generally accepted set of core principles, underpinned by philosophy, methodological work and expert opinion. A considerable amount of work by leading review methodologists has been undertaken in developing guidance in the design and conduct of systematic reviews. Such guidance has been codified to enable researchers to judge whether a given systematic review is likely to be of high or low quality.

Key features of systematic reviews

1. A transparent, comprehensive search strategy.
2. Clear pre-specified inclusion/exclusion criteria.
3. Explicit methods for coding, quality appraising and synthesising included studies.

Quality of systematic reviews

Systematic reviews, like any other form of research, can be of variable quality. To ensure the highest quality in design and methods in undertaking a systematic review methodologists have developed a number of guidance statements.

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement (Moher et al., 1999; Shea et al., 2001) (which supersedes the QUOROM Statement) is a minimum set of items for reporting systematic reviews and meta-analyses, developed through methodological work. The aim of the PRISMA Statement is to help authors improve the reporting of their systematic reviews and meta-analyses. PRISMA focuses on systematic reviews of randomised controlled trials but it can also be used as a basis for reporting systematic reviews of other types of research, particularly evaluations of interventions. PRISMA may also be useful for critical appraisal of published systematic reviews. The PRISMA Statement consists of a 27-item checklist and a four-stage flow diagram. The 27 items are included under seven subsections: title, abstract, introduction, methods, results, discussion, funding. So, for each stage of the systematic review, explicit guidance is given on how it should be reported. Authors of systematic reviews in any field, including education, are recommended to use the PRISMA checklist in the design, conduct and reporting of their reviews. The PRISMA flow diagram depicts the flow of information through the four different phases of a systematic review, including the number of records identified in the searches, the number assessed for eligibility, inclusion and exclusion, and the reasons for the exclusions.

Stages of a systematic review

A systematic review can be seen as having seven main stages which are well established in health care, education and social science research:

1. *Research question.* Development of a well focused, clear research question which can be addressed by a systematic review; establishing the review team and the parameters of the review.
2. *Protocol.* Development of a protocol or plan of the review, including an *a priori* statement of the design and methods for each stage of the review.
3. *Information retrieval and study selection.* Development of a search strategy, searching and screening to identify/select the studies included in the review.
4. *Coding.* Extraction of data from each of the included studies using a coding form developed for the review.
5. *Quality appraisal.* Assessment of risk of bias in each of the included studies using, for example, a tool developed from the CONSORT Statement (see below).
6. *Synthesis.* Results of all the included studies are combined (this may include a meta-analysis).
7. *Report writing.* The systematic review is disseminated through a report or published article.

Detailed guidance on methods for undertaking a systematic review

In the following, detailed guidance on methods for undertaking the seven stages of a systematic review have been applied to an exemplar review in the field of educational research.

Exemplar: Writing review

The title for the exemplar review is: 'A systematic review of the effectiveness of writing interventions on written composition' (hereafter 'writing review'). This is an effectiveness review which means that the primary studies included in the review would be studies using an experimental design.

1. Research question

The research question is the first stage in any systematic review. Once a question is established which is of substantive, methodological or policy importance, a rapid scope (preliminary search of the main electronic databases) can be undertaken to check whether any previous systematic reviews have already been undertaken. If no previous reviews exist or if any previous reviews are out of date, a systematic review is justified and the parameters of the review can be established. The parameters limit the scope of the review, and include such items as the language and publication dates of the studies to be included. All parameters require a justification. The research question can be framed in terms of the participants, interventions, outcomes, study designs (PICOS) categories (Moher et al., 1999 – see below).

Once it has been established that a systematic review is justified, the review team can be set up. An important aspect of the design of a systematic review is that it should be undertaken by a team of researchers rather than by a single researcher. This is because more than one person is required to assure the conduct of the review is of the highest quality. For example, double screening and coding (data extraction) are recommended to ensure that there is minimal error or bias in the review. Both substantive and methodological experts are required to work as a team to develop the research question, develop the search strategy and interpret the findings of the review; methodological experts are required to quality assure the review and statistical experts are necessary to undertake any meta-analysis.

Summary

- The research question should be an important substantive, methodological or policy question.
- It establishes the parameters and restricts the scope of the review; it drives all the subsequent stages of the review.
- The research question can be framed in terms of the PICOS categories.

2. Protocol

The protocol or plan for the research describes the design and methods of the systematic review in advance of the identification of the studies included in the review. The design of the review will include such features as the conceptual underpinning of the review, the parameters of the review and the rationale for the research question being addressed by the studies included in the review. In addition, the protocol specifies the study characteristics,

using the categories participants, interventions, outcomes, study designs (PICOS) (Moher et al., 1999) and these will be used as criteria for eligibility for inclusion in the review. The research question and objectives, the scope of the review, its parameters and strategy for information retrieval, inclusion/exclusion criteria, methods for searching, coding (data extraction) and the development of an assessment of risk of bias tool for quality appraisal of included studies are all pre-stated in the protocol. This will reduce the possibility of reviewer selection bias and inclusion bias.

The main reason for developing a protocol in advance of undertaking the review is to limit bias potentially introduced by the reviewers. If the main research question is developed and the methods are specified in advance *before* the literature is identified, this prevents the research question being altered by the data. For example, if it is pre-specified that only randomised controlled trials (RCTs) will be included but during the search a large quasi-experimental study is found the results of which support a prior hypothesis, including this study at this stage would require a change to the methods of the review and this may introduce a potential source of bias. While the review might refer to the quasi-experiment to set the experimental studies in context, the main finding should, in this case, be based on a synthesis of the experimental studies. Reviewers may go on and develop a further review protocol that states that quasi-experiments will be included in an update of the review but they should not be included in the current review as they were not pre-specified.

Although the process of undertaking a systematic review can include an iterative process, any changes after the finalising of the protocol have to be made explicit and justified, and for this reason the protocol is generally sent for peer review and 'published' in a public place (website or online journal) to increase the rigour and transparency of the review.

The background to the review will include its rationale in the context of what is already known, previous theoretical and empirical research including previous systematic reviews, informed by a 'rapid scope' of the literature and the policy and practice context to the review. The protocol also includes the parameters of the review, the inclusion criteria (with justifications), the categories for coding and the criteria for assessing risk of bias in the included studies. The proposed nature of the synthesis is also pre-specified in the protocol. In Figure 30.1 a brief exemplar protocol for the writing review is presented.

Research question: What is the effectiveness of writing interventions on the written compositional skills of children aged 7 to 16 in mainstream school settings?
Objective: The objective of the review is to systematically search for, identify, locate and quantitatively synthesize (meta-analyse) the high-quality evidence of the effectiveness of writing interventions aimed at either improving or reducing or preventing writing difficulties of children and young people aged between 7 and 16 in mainstream school settings.
Rationale for review/background: Confidence and accuracy in written expression should be an attainable outcome for all children in mainstream education. In addition, quality of written expression is related to children's ability to access and achieve in all areas of the curriculum in both the primary and secondary phases of education. A number of interventions have been developed for those target groups which researchers have identified as underachieving at writing. Although there have been a number of systematic reviews and meta-analyses in the topic area of writing, a tertiary review identified no review that synthesised the experimental research on the effectiveness of writing interventions in all writing genres (Torgerson, 2007). There is therefore a need for such a review to be undertaken in

order to inform policy and practice.

Conceptual issues: The conceptual issues include the nature of writing development in a variety of genres, theories of the development of writing abilities, writing interventions and outcomes, potential mediators and moderators, and conceptual issues regarding appropriate research designs to address an effectiveness question.

Design and method: The design of the review is a full systematic review; design and methods of the review are informed by the Campbell Collaboration policy briefs (see http://www.campbellcollaboration.org/); 'Systematic reviews: CRD's guidance for undertaking reviews in health care' (see http://www.york.ac.uk/inst/crd); the 'Cochrane Collaboration Handbook' (see http://www.cochrane-handbook.org/); the (1994) *Handbook of Research Synthesis* (eds) Cooper, H, Hedges, L. and Torgerson, C. (2003) *Systematic Reviews.* London: Continuum.

Design of studies included: Studies that can adequately address the research question (which is an effectiveness question) are high-quality evaluations of interventions to improve the quality of expression in pupils' written work using experimental designs: randomised controlled trials and quasi-experiments. This is because, in order to establish causality (i.e. to be able to state that a specific teaching practice *causes* an improvement in written outcomes) study designs which can adequately control for all other known and unknown variables that could affect outcome are required (Cook and Campbell, 1979; Shadish et al., 2002). The review will focus on research evidence from academic journals and other published research and, in order to limit the possibility of publication bias, research from the difficult-to-locate 'grey' literature:

1. Randomised controlled trials (allocated at either the individual level or cluster level e.g. class/school/district).
2. Quasi-experimental studies of any design (including regression discontinuity design, interrupted time series design).

Studies in which at least one of the groups received a writing intervention compared to standard practice ('business-as-usual') or an alternative writing intervention will be included. Studies in which the control group did not receive any writing instruction will be excluded. Citation searches will be conducted on any located previous systematic reviews/ meta-analyses.

Types of participants in included studies: Studies in which participants have English as a first, second or additional language will be included. Studies evaluating interventions in children or young people aged 7 to 16 years in a full-time mainstream educational setting will be included. Studies evaluating interventions in children of all learner characteristics attending mainstream schools and classes will be included.

Types of interventions (and comparisons) included: Studies evaluating any whole text writing intervention will be included, for example: provision of model writing structures; guided practice; advanced planning strategies (strategies for developing, evaluating and organising ideas); collaborative or cooperative editing and revision; self-regulated strategy development (e.g. goal setting, self-monitoring, self-regulation), strategies for editing and revision, text analysis, writing prompts, strategies for composing, editing and revising different text types, strategies for directing processes for planning and composing. Strategies for improving writing in the following genres will be included: descriptive writing, expository writing, narrative writing, poetry, drama, instructional writing, argumentation, letter writing, discursive writing and persuasive writing.

Types of outcomes included: Studies will be included if they contain at least one of the following kinds of quantified outcomes: holistic writing quality, length of composition, planning and composing times, essay elements, essay

coherence, maturity of vocabulary, reader sensitivity, productivity, elements or features of writing in different genres, e.g. quality of argument, quality of persuasiveness, quality of description, quality of narrative writing, text structure.

Proposed codings for assessment of risk of bias in included studies: A modified version of the CONSORT Guidelines will be used in order to develop a tool to assess the risk of bias in the included randomised and quasi-experimental studies. This assessment of methodological quality of the included studies will include reviewer judgement of the following: method to generate allocation to groups and concealment of that allocation; evidence of sample size calculation; eligibility criteria specified; blinding of intervention provider, participants and outcome assessor; presentation of estimate of effect size and its precision; attrition; primary analysis, i.e. intention-to-treat or on-treatment analysis. A subgroup analysis of the higher quality trials will be undertaken, if appropriate.

Methodsfor coding (extracting datafrom) included studies: Data from the included studies will be extracted onto a specially designed coding form. Data to be extracted will include: country, setting, aims and objectives, research design, participants, inclusion criteria, interventions and control or comparison conditions, outcomes, results.

Synthesis: A narrative synthesis will be undertaken to combine the results of the included studies and, if appropriate a meta-analysis (statistical synthesis) will be undertaken.

Proposed quality assurance procedures: Independent double screening, data extraction, quality appraisal (assessment of risk of bias) and extraction of quantified outcomes will be undertaken. Procedures to assure the quality of each stage of the review will be set up.

References

Cook, T. D. and Campbell, D. (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Boston, MA: Houghton-Mifflin.

Shadish, W. R., Cook, T. D. and Campbell, T. D. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton-Mifflin.

Torgerson, C. (2007) 'The quality of systematic reviews of effectiveness in literacy learning in English: a "tertiary" review', *Journal of Research in Reading,* 30(2).

Figure 30.1 Exemplar protocol

Inclusion criteria

(1) Topic: Studies about writing in English-speaking countries (English as *first, second or additional language).*

(2) Study design: Studies with designs where there is a control or comparison group - randomised controlled trials (individual or cluster); quasi-experiments (case control studies, cohort studies, regression discontinuity studies, interrupted time series).

(3) Participants: Studies where the participants are aged between 7 and 16 years (inclusive) and in full-time mainstream education.

(4) Interventions: Studies evaluating whole-text writing interventions in the following genres: description, expository writing, narrative writing, poetry, drama, instructional writing, writing argument, letter writing, discursive writing and persuasive writing.

(5) Intervention and control or comparison treatments: Studies in which at least one of the groups received a

writing intervention compared to standard practice ('business-as-usual'), or an alternative writing intervention.
(6) Outcome: Studies in which participants are measured at post-test on a writing outcome, e.g., holistic writing quality, length of composition, planning and composing times, essay elements, essay coherence, maturity of vocabulary, reader sensitivity, productivity, elements or features of writing in different genres, e.g. quality of argument, quality of persuasiveness, quality of description, quality of narrative writing, text structure.
Exclusion criteria
(1) Topic: Studies about writing English as a *foreign language.* **(2) Study design:** Studies with designs where there is no control or comparison group. **(3) Participants:** Studies in which the participants are below the age of 7 or above the age of 16 or in which the participants do not attend mainstream schools. **(4) Interventions:** Studies which do not evaluate whole-text interventions in the stated writing genres. **(5) Intervention and control or comparison treatments:** Studies in which the control group did not receive any writing instruction. **(6) Outcome:** Studies in which participants are not measured at post-test on a writing outcome.

Figure 30.2 Exemplar inclusion and exclusion criteria

The inclusion and exclusion criteria are developed alongside the protocol and are based on the research question and parameters of the review. The inclusion and exclusion criteria are used for checking all studies that could be potentially included in the review in order to determine eligibility for inclusion. In Figure 30.2 the inclusion and exclusion criteria for the exemplar writing review are presented. The criteria focus on the topic area, the study design, the participants, the interventions, the comparison or control conditions and the outcomes. Each inclusion criterion is mirrored by an exclusion criterion, which enables the process of screening to be operationalised and the reasons for exclusion to be documented.

Summary

- The protocol is a plan of the review, written in advance of study identification or selection in order to limit bias.
- It contains the design and methods of the review, including the research question, search strategy, inclusion/exclusion criteria and proposed methods for synthesis.
- The protocol can be changed during the course of the review but all changes should be documented and justified.

3. Information retrieval and study selection

Information retrieval and study selection refers to the methods for searching, locating and checking the inclusion eligibility of potentially relevant studies. These methods should be rigorous to ensure that a high proportion of the eligible published and unpublished studies will be located, retrieved and included. Systematic information retrieval is critical to a systematic review as it ensures an unbiased compilation of potentially relevant research by minimising bias and maximising coverage. The main thrust of the search is likely to use the electronic sources, although other

methods of retrieval can supplement the electronic searches. For the electronic searches, an exhaustive search strategy is important. High sensitivity, that is identifying as many relevant papers as possible, may result in low precision (i.e. most papers identified are not relevant to the review), so a judicious balance between the two is recommended. Ideally, an experienced information scientist should be consulted for this aspect of the review.

The strategy used to search the electronic databases is usually based on one or more of the PICOS elements used to produce the inclusion and exclusion criteria (see above). The information contained in database abstracts is limited and rarely reports all of the inclusion criteria required by the review. For this reason it is advisable to use as few PICOS elements as possible, to avoid missing relevant material. The search strategy constructed for the exemplar writing review used only two facets, intervention (writing interventions) and outcome (writing composition). Each facet should contain a variety of terms to capture that element of the review question. Terminology will differ from article to article, even when the same topic is being addressed, so it is important to use synonyms to capture as many relevant papers as possible. The use of indexing terms can help with this, so where they are available they should be used alongside free text (or natural language) terms.

The creation of the search strategy is an iterative process and the strategy may go through a number of versions in response to feedback on the material retrieved. Ideally, the search strategy should be peer-reviewed, and it may evolve still further in the light of this process. Copies of all search strategies should be kept, along with information about which databases were searched and when the searching was undertaken. This will assist with the writing of the final report and enable critical appraisal of the search element of the review. Once the basic search strategy for the electronic databases has been finalised, the searches can be undertaken.

As mentioned above, the main thrust of identifying research studies is likely to be on the electronic searches, but these can be supplemented through hand searching of key journals. This may be of particular benefit if the area is a 'niche' subject with a few specialist journals publishing relevant material or if the subject is difficult in terms of key words that can be used to identify relevant material. Furthermore, older publications, in particular, may not specify their design particularly well, making some relevant studies difficult or impossible to locate using electronic means. Relevant studies can also be identified through citation searching and checking the bibliographies of previous systematic reviews and seminal studies. Also, reviewers may contact authors of relevant publications to ask them if they are aware of any other relevant studies, including their own, particularly unpublished studies. Nevertheless, however rigorous the method of searching, a number of potential sources of bias can be introduced through the search. These include publication bias, language bias, time lag bias and database bias, all of which have the potential to introduce a source of bias into the review.

Publication bias is the phenomenon whereby studies with 'positive' findings are more likely to be reported in the peer-reviewed literature than studies with null or negative effects. If primary research studies remain unpublished and if there is a relationship between non-publication and their outcomes this can, in turn, affect the findings of systematic reviews. In systematic reviews potential sources of publication bias are not searching the grey literature for unpublished (but in the public domain) reports and studies not having been published (therefore not able to be included in the review). Publication bias has been described as the Achilles' heel of any literature review (Torgerson, 2006). If non-publication was a random event this would only matter in our level of uncertainty. A meta-analysis that indicates a non-statistically significant benefit of an intervention may, in fact, be recording a Type II error, i.e.

concluding erroneously there is no statistically significant difference, when in fact, if all of the studies ever undertaken had been assembled, then the difference would have been statistically significant. However, this is the lesser of the two problems: the second problem of bias is more serious. Usually there is a reason why studies are not published and this often relates to the study's findings. A study that finds either no effect or a difference going in the opposite direction to that hypothesised has a lower possibility of being published. Authors of such studies may feel journals are less likely to accept such studies (a self-fulfilling prophecy) and not submit them, while editors and reviewers may be more likely to reject them. Even when negative studies are published the process often takes longer than for positive studies. In contrast, studies that have a positive result, especially a statistically significant one, are often fast-tracked by the authors for submission and are more likely to be accepted by referees and editors than negative studies. Consequently, at any one time the published literature is more likely to be over-representative of positive results than negative findings.

The other biases that can affect systematic reviews include language bias, time lag bias and database bias. It may be the case that important papers are not published in English and are excluded from the review because of the cost of requiring necessary translations. Time lag bias is a form of publication bias described above whereby 'negative' studies take longer to publish than positive studies. Database bias may occur if the choice of databases means that a significant proportion of unpublished (or 'grey' literature) studies are excluded because they are not present in the narrow choice of databases used for the search.

Search strategy for ERIC [Dialog DataStar):
1. writing-composition.de.
2. writing-processes.de.
3. (descriptive near (write or writing or written or essay or essays or composition or paper or papers or text or texts or assignment or assignments or document or documents or prose)).ti,ab.
4. (discursiv$ near (write or writing or written or essay or essays or composition or paper or papers or text or texts or assignment or document or documents or prose)).ti,ab.
5. (written adj expression).ti,ab.
6. 1 or 2 or 3 or 4 or 5 **[brings together all the terms for writing composition]**
7. (model$ near (write or writing or written) near structure$).ti,ab.
8. collaborat$ near (edit or editing or revision$)
9. ((strategy or strategies) near (edit or editing or edits or compose or composition or revise or revision or write or written or writing)).ti,ab.
10. 7 or 8 or 9 **[brings together all the terms for writing interventions]**
11. 6 and 10 **[retrieves papers that include terms for both a writing intervention and a writing outcome]**

Key:

.de. = restrict search to index term

.ti,ab. = restrict search to title or abstract

$ = truncate term

adj = finds terms next to each other

near = finds terms within five words of each other

Figure 30.3 Exemplar search strategy

Details about how the search strategy was developed and which databases and journals will be searched should be clearly described in order to permit scrutiny and replication. To ensure that any search is replicable the PRISMA Statement recommends that all information sources are described in detail, including the date the searches were undertaken. It also recommends that at least one full electronic search strategy is presented in the report, together with any limits in order that it could, in theory, be replicated. A simplified version of a search strategy for the exemplar writing review is reproduced in Figure 30.3. (This is for illustrative purposes only – the original contained many more search terms.)

During screening the inclusion criteria are applied to the results of the searches in a three-step process which should be pre-specified in the protocol: prescreening (to filter out studies that are immediately and easily identified as being irrelevant to the review), first-stage screening (of titles and abstracts) and second-stage screening (of full papers). The identified articles are checked against the predetermined criteria for eligibility and relevance. This process should be undertaken rigorously and quality assured in order to minimise bias. It is recommended that two reviewers screen at each stage independently and then compare their decisions. If this is not possible due to resource constraints the database of potentially relevant studies can be screened by one reviewer, with a random sample of studies screened by a second reviewer at each stage. If this process is adopted the inter-rate reliability of the screening of the two reviewers should be checked through, for example, the calculation of a Cohen's Kappa statistic, and if agreement is low (as demonstrated by a low Cohen's Kappa statistic), then the entire dataset will need to be screened by two reviewers (independent double screening).

Summary

- The research question determines the limits of the search which should be comprehensive, explicit and replicable.
- The main focus of the search strategy is likely to be on the electronic searches, but these should be supplemented by other means, such as hand searching of key journals or citation searching.
- Studies should be screened for selection into the review using pre-specified inclusion and exclusion criteria.

4. Coding

Once the screening has been completed the included studies should be coded. Coding, or data extraction, is the process by which the included studies are described and classified. A paper-based or electronic coding instrument should list all the items for which data will be sought. This will include, as a minimum, data extraction of information about the bibliographic details of the study and its aims and objectives, and key items using the PICOS categories (see above): participants, intervention, control or comparison conditions, outcomes and study designs. It will also

state the quantified outcomes which will be extracted (e.g. the means and standard deviations of all pre- and post-tests for all groups).

In addition, the items which will be used to appraise the quality of the included studies (see below) should be coded, and the quality assurance procedures for ensuring the reliability of the coding should be recorded. Ideally this should involve independent double coding, with a plan for comparing information on the coding form and procedures to follow when two reviewers disagree. The coding instrument should ideally be piloted using draft versions to extract data from a dozen or so papers to test the efficacy of the instrument with a relatively small sample of studies: it can then be amended if necessary. If possible, the coding should be undertaken 'blind' to authors of the included studies, although to do this may be costly and time-consuming.

5. Quality appraisal

In order to limit the potential for introducing bias into a systematic review because of design bias it is necessary to critically appraise the included studies in order to assess the potential for risk of bias. In the case of randomised controlled trials, which vary in quality, pooling the results of a number of RCTs with risk of bias due to methodological shortcomings in their design needs to be explored in the review. Methodological work, mainly in the field of healthcare research, has

led to the development of a number of tools designed to quality appraise RCTs. The Consolidated Standards of Reporting Trials (CONSORT) Statement (Schultz et al., 2010) is not a quality assessment tool, but a minimum set of recommendations for reporting RCTs in a standardised way in order to increase transparency and to help with critical appraisal of trials. A form of the CONSORT Statement, adapted to make it relevant to educational research, can be used to develop a quality appraisal (or assessment of risk of bias) tool. The CONSORT Statement comprises a 25-item checklist and a four-stage flow diagram. The tool should focus on the most important reporting issues such as trial design and analysis, for example whether the allocation was independent and concealed, whether a sample size calculation was undertaken, whether the unit of analysis matched the unit of allocation, and whether there was high or differential attrition between the arms of the trial. The four-stage flow diagram depicts the flow of participants through the key stages of the trial: enrolment, allocation, follow-up and analysis. Quality appraisal of studies should include whether or not a CONSORT-type flow diagram was included in the report. This is important as a significant source of bias can be introduced through high attrition and the flow diagram is a visual way of presenting these data. The potential impact of the quality of the included studies should be considered in the synthesis.

Study characteristics (e.g. publication year)
Methodological characteristics (e.g. method of assignment to condition, study design)
Participant characteristics (e.g. gender, SES, baseline writing)
Intervention characteristics and implementation fidelity
Control/comparison characteristics and implementation fidelity
Outcome measures

Figure 30.4 Exemplar coding book

Summary

- Coding involves extraction of information from the included studies to describe and classify the studies.
- The kind of data extracted will depend on the research question and the type of synthesis that will be undertaken.
- The categories for coding should be pre-specified, piloted on a sample of studies and based on the PICOS categories.
- Quality appraisal involves assessing the included studies for key sources of risk of bias and methodological rigour.
- It can be undertaken using a tool developed from the CONSORT Statement.
- Methods for quality assuring the coding and quality appraisal should be predetermined, and rigorously undertaken and reported.

6. Synthesis

The synthesis involves combining the results of the individual studies using a framework or structure. This can be done in a number of different ways. For example, a 'qualitative' or 'narrative' synthesis may be undertaken, or a quantitative or statistical 'meta-analysis' may be used to combine the results of homogeneous studies and increase the power and precision in the measurement of effect sizes.

In a 'narrative' synthesis the included studies are grouped thematically in terms of their characteristics; for example different varieties of the intervention being evaluated and different learner characteristics, and then commonalities between studies (whether of all the studies or a sub-group of them) can be described. However, systematic reviews of randomised controlled trials and quasi-experiments often, but not always, use statistical techniques (meta-analysis) to combine quantitatively the results of the eligible studies. Whether a narrative or a statistical synthesis is undertaken, the strengths and limitations of each of the included studies (based on the quality appraisal undertaken in step 5 –see above) should be taken into account in drawing conclusions based on the results of the synthesis.

Meta-analysis

Probably the most frequently used method of synthesising quantified data is meta-analysis. Simply put, a meta-analysis combines all the studies to give an overall or summary estimate of effect. Generally, meta-analyses use a process of giving greater weighting to larger studies as these are usually likely to be the most reliable studies in the review. Authors of the systematic review should consider, first, whether a meta-analysis is appropriate, given that in any systematic review the included studies may not be homogeneous in terms of participants, interventions and outcomes, etc. If a meta-analysis is deemed to be appropriate, authors should include, as a minimum, the following: a pooled effect size of all studies eligible to be included in the meta-analysis, with confidence intervals; an indication of how heterogeneity between the studies was explored; whether a fixed-effect model or a random-effects model was used for the meta-analysis; and the pre-specified sub-group and sensitivity analyses that were undertaken. An indication of whether and how the potential for publication bias has affected the results of the meta-analysis should be included, for example through the use of a 'funnel plot'. Given sufficient numbers of studies, a meta-regression

analysis can be undertaken to explain some of the heterogeneity observed within the component studies. For example, in the writing review whether or not age of pupil affects outcomes or whether underlying ability interacts with the intervention could be explored through a meta-regression. Furthermore, it is possible to see if the results are affected by the underlying quality of the component studies. Do methodologically weak studies, for example, generate larger effect sizes? The advantage of a meta-regression is that it may help to explain the heterogeneity of findings. However, there does need to be caution when interpreting any findings from a meta-regression. First, the statistical power of any meta-regression is relatively small; consequently there may be relatively important interactions that the analysis does not uncover. Second, some false positive interactions may be observed. Despite there being a statistically significant interaction with, say, gender, in truth no such relationship may exist. Therefore such findings should be used to generate hypotheses and be confirmed, ideally, in a large robustly designed RCT.

There are a number of different techniques that can be used to undertake a meta-analysis, a detailed description of which is beyond the scope of this chapter (see Chapter 46 in this book for detailed guidance on the use of meta-analytical techniques; see also Lipsey and Wilson, 2001 for a detailed text on meta-analysis).

Summary

- The synthesis involves the combining of the results of the review.
- Where possible, the nature of the synthesis should be proposed in advance.
- It can take a variety of forms, including a 'narrative' synthesis or a meta-analysis.
- The quality appraisal judgements for each included study should be taken into consideration in the synthesis.
- Meta-analyses are sometimes appropriate in systematic reviews of experimental research in order to provide a pooled effect size of a group of homogeneous studies.
- If a meta-analysis is undertaken, the possible presence of publication bias should be investigated.

7. Report writing

Like all primary research, systematic reviews should be written up and published as soon as possible after their completion. Indeed, there is, arguably, a stronger imperative for publishing reviews as swiftly as possible as they will tend to become outdated more quickly than primary research as the search strategy for the systematic review is time limited. The process of writing up the final report ought to be guided by the PRISMA Statement to ensure its results are deemed to be of high quality. The report should refer to the existing protocol and be written so that it is as accessible as possible to the widest audience, including policy-makers and practitioners.

Summary

- The design, methodology and methods of a systematic review should be written up in a report and published and disseminated widely in a timely manner, as reviews soon become outdated.
- The process of writing the report can be guided by the PRISMA Statement which will increase the rigour of its reporting.
- The report should be accessible to the widest possible audience, including policy-makers, practitioners and

researchers.

Conclusion

With the explosion of research endeavour across the world it is difficult or impossible for researchers, practitioners and policy-makers to keep abreast of new research findings by reading all of the primary research. Systematic reviews enable researchers and others to access the literature in a comprehensive and unbiased manner. In this chapter the basic design and methods underpinning a systematic review have been described. Systematic reviews are an essential precursor to sophisticated synthesis methods such as meta-analysis. A meta-analysis is only as good as its component studies. Should these be from a biased sample of studies then even the most sophisticated statistical techniques cannot rescue the results. Consequently meta-analysts should pay careful attention to how the studies they are including in their statistical synthesis were identified.

In summary, systematic reviews are scientific reviews which have explicit, transparent and theoretically replicable designs and methods. The key features of their design limit any potential biases. They can increase evidence-based education by synthesising a body of literature in a topic area to address a specific research question. They can also identify 'gaps' in the literature, inform the design of a randomised controlled trial or a future research agenda. However, systematic reviews, like all forms of research, can vary in terms of their quality. Therefore any users of systematic reviews should pay critical attention to the quality of their design, conduct and reporting.

Questions for further investigation

1. Find a systematic review in your area of interest. Prepare a critique of the review in terms of transparency, explicitness and replicability. Make an assessment of the reliability of the findings of the review.
2. Write a research question for a systematic review in a topic area of interest together with a compelling rationale for its significance. Develop the following: parameters for the review; inclusion and exclusion criteria; methods for searching, screening, coding and synthesis. Note any challenges to designing this review.

Suggested further reading

Campbell Collaboration at: http://www.campbellcollaboration. org/. The Campbell Collaboration prepares, maintains and disseminates systematic reviews in education, crime and justice and social welfare.

Cochrane Collaboration at: http://www.cochrane-handbook. org/. The Cochrane Collaboration prepares, updates and promotes systematic reviews in health care. See also: Higgins, J. P. T. and Green, S. (eds) *Cochrane Handbook for Systematic Reviews of Interventions,* Version 5.0.2 (updated September 2009).

Meta-analysis in Education Research at: http://www.dur. ac.uk/education/meta-ed/. The Economic and Social Research Council (ESRC) Researcher Development Initiative (RDI): Training in Quantitative Synthesis

(Meta-analysis) aims to develop understanding of the design, methodology and methods of meta-analysis.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., PRISMA Group (2009) 'Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement', *BMJ,* 339.

References

Chalmers, I., Hedges, L. V. and Cooper, H. (2002) 'A brief history of research synthesis', *Evaluation and the Health Professions,* 25: 12-37.

Cooper, H. and Hedges, L.V. (eds) (1994) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

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

Glass, G. V., McGaw, B. and Smith, M. L (1981) *Meta-analysis in Social Research.* Beverly Hills, CA: Sage.

Lipsey, M. W. and Wilson, D. B. (2001) *Practical Meta-analysis,* Applied Social Research Methods Series 49. London: Sage.

Moher, D., Cool, D. J., Eastwood, S., Olkin, I., Rennie, D., Stroup, D. F. (1999) 'Improving the quality of reports of metaanalyses of randomized controlled trials: the QUOROM Statement', *Lancet,* 354: 1896-900.

Mulrow, C. (1994) 'Rationale for systematic reviews', *BMJ,* 309: 597.

Petticrew, M. (2001) 'Systematic reviews from astronomy to zoology: myths and misconceptions', BMJ, 322: 98.

Schultz et al. (2010) CONSULT Statement (online). Available at: http://www.consort-statement.org/ (accessed 20 September 2011).

Shea, B., Dube, C. and Moher, D. (2001) 'Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools', in Egger, M., Davey-Smith, G. and Altman, D. (eds) *Systematic Reviews in Healthcare: Metaanalysis in Context* (2nd edn). London: BMJ Publishing Group.

Torgerson, C. (2003) *Systematic Reviews.* London: Continuum. Torgerson, C. (2006) 'Publication bias: the Achilles' heel of systematic reviews?', *British Journal of Educational Studies,* 54(1).

Torgerson, D. and Torgerson, C. (2008) *Designing Randomised Trials in Health, Education and the Social Sciences.* London: Palgrave Macmillan.

How can we get solid results from literature: The use of systematic review design and methods

Professor Carole Torgerson
Durham University, UK
carole.torgerson@durham.ac.uk

Professor David Torgerson
University of York, UK
david.torgerson@york.ac.uk

Milan March 15th 2016

Introduction

Systematic review

Design

- Synthesis and critical evaluation of existing research in a topic area
- Secondary research

History

- Long history in a number of research areas (e.g., astronomy, health care, criminal justice) as well as in education research

What is a systematic review?

- A systematic review
  - is a synthesis of all the studies relevant to address a specific research question
  - has explicit, transparent, replicable methods
  - states in advance the criteria for including studies
  - searches exhaustively for all the relevant studies within a pre-defined area
  - negative and positive studies included to give an overall impartial view of the field
  - narrative synthesis or meta-analysis

Rationale for systematic reviews

- Evidence-based education and social science
- Challenge of keeping up with the enormous volume of research being produced
- Not all study results get into the public domain therefore rigorous and systematic searches are required to ensure all the relevant evidence is considered
- Research evidence is variable in quality and can be biased
- Reduction in unnecessary research
- Identify gaps in the evidence to identify areas where research is needed to inform decisions, guidelines, policy

Key features of a systematic review

- Explicit, transparent, replicable search
- Critical evaluation (quality appraisal) of all included studies
- Synthesis: narrative or quantified (meta-analysis)

Key stages in systematic reviewing

- Key review question (and conceptual framework)
- Search strategy
- Inclusion/exclusion criteria
- Coding and mapping (initial organization of data)
- In-depth review (identifying and exploring patterns in the data)
- Techniques for systematic synthesis (integration of the data – narrative or meta-analytic)

What is the question?

What data are available?

What patterns are in the data?

How robust is the synthesis?

What are the results?

Protocol

- The protocol is developed a priori to establish:
  - the research question
  - the methods for conducting the review
  - inclusion/exclusion criteria
  - data extraction procedures
  - critical evaluation (quality appraisal) of included studies

Inclusion criteria

- Although all relevant literature should be identified not all will be included in the review
- Literature may be excluded if:
  - not directly relevant or not primary research (e.g., an editorial)
  - low quality (see below)
  - review is restricted to certain study types (e.g., only randomised controlled trials)

Methodological quality

- What about 'low quality' studies?
  - All studies are likely to have weaknesses (methodological quality is on a range or continuum)
  - Exclusivity restricts the scope and scale of the analysis and generalizability
  - Inclusivity may weaken confidence in the findings
  - Some methodological quality is in the "eye-of-the-beholder"
  - to the key research question

Synthesis

- Narrative synthesis
- Meta-analysis combines 2 or more similar studies to provide more precise estimate of effect
- Effect size is usually calculated by dividing the difference in mean post-test scores by the standard deviation of the control group or by a 'pooled' standard deviation

Glass *Educational Researcher* 1976;5:3-8.

Example of systematic review: Driver education

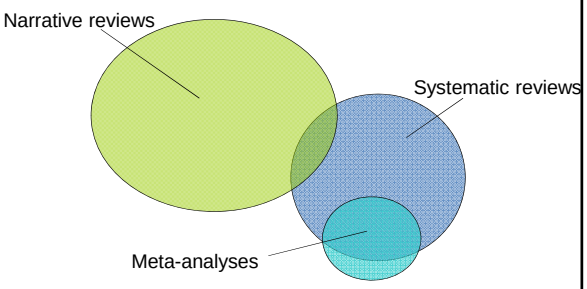
- 2000 UK government launched 10-yr. plan to reduce road deaths associated with young motorists (drivers aged 17-21 7% of all drivers but involved in 13% of injurious road accidents). Part of this plan included the introduction of driver education in schools.
- Systematic review found 3 studies with experimental designs showing an acceleration of licence acquisition (risk factor for accidents) but NO reduction in RTAs (actually a slight increase).
- In contrast, a non-systematic review showed driver education to be advantageous.

Achara et al. Lancet 2001; 358:230

Types of systematic reviews

- Scoping review
  - Broad mapping of the literature in a field
- Systematic review
  - In depth review relating to specific review question (may or may not include meta-analysis; may be used to address substantive or methodological question, for synthesis or hypothesis generation)
- Tertiary review
  - Overview of systematic reviews in a field

Literature reviews - conceptual relations



Traditional review (narrative review or 'expert' review)	Systematic review
Research question often not explicit	Explicit research question
No protocol	Protocol (or plan) developed (published) in advance of undertaking review
Criteria for including and excluding studies not explicit	Pre-specified explicit criteria for including and excluding studies
Arbitrary, biased selection of literature	Systematic, comprehensive selection of literature
Data from included studies variably emphasised depending on reviewer's perspective or argument	Data from included studies extracted in a pre-specified, systematic way Explicit 'weighting' of evidence from included studies

Traditional review	Systematic review
Potential for bias in the studies either not considered or only considered for some studies	Systematic critical appraisal of all the studies to uncover potential for bias
Usually sole reviewer or if more than one reviewers no systematic procedures	Minimisation of bias by systematic procedures by more than one reviewer
Review procedures not replicable	Review procedures replicable
'Subjective' – therefore potential for bias to be introduced at every stage of review	'Objective' – therefore potential for bias minimised at every stage of review

Advantages of SR

- Use explicit, replicable methods to identify relevant studies
- Use established or transparent techniques to analyse those studies
- Aim is to limit bias in the identification, and evaluation of studies and in the integration or synthesis of information applicable to a specific research question.

Terminology

Systematic review

- A rigorous way of finding, selecting, evaluating and collating all the available research evidence to ask a specific question

Meta-analysis

- A statistical technique used to combine the results of two or more studies into a single combined quantitative estimate

Origins

1952: Hans J. Eysenck concluded that there were no favorable effects of psychotherapy, starting a raging debate which 25 years of evaluation research and hundreds of studies failed to resolve

1978: To prove Eysenck wrong, Gene V. Glass statistically aggregated the findings of 375 psychotherapy outcome studies

Glass (and colleague Smith) concluded that psychotherapy did indeed work "the typical therapy trial raised the treatment group to a level about two-thirds of a standard deviation on average above untreated controls; the average person received therapy finished the experiment in a position that exceeded the 75th percentile in the control group on whatever outcome measure happened to be taken" (Glass, 2000).

Glass called the method "meta-analysis"

(adapted from Lipsey & Wilson, 2001)

Historical background

- Underpinning ideas can be identified earlier:
  - K. Pearson (1904)

    Averaged correlations for typhoid mortality after inoculation across 5 samples
  - R. A. Fisher (1944)

    "When a number of quite independent tests of significance have been made … although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance" (p. 99).

    Source of the idea of cumulating probability values
  - W. G. Cochran (1953)

    Discusses a method of averaging means across independent studies

    Set out much of the statistical foundation for meta-analysis (e.g., inverse variance weighting and homogeneity testing)

(adapted from Lipsey & Wilson, 2001 and Hedges, 1984)

Cochrane and Campbell Collaborations

- Cochrane Collaboration
http://www.cochrane.org/

- Campbell Collaboration
http://www.campbellcollaboration.org/

Some recent findings from meta-analysis in education

Bernard *et al.* 2004
- Distance education and classroom instruction - 232 studies, 688 effects - wide range of effects ('heterogeneity'); asynchronous DE more effective than synchronous

Pearson *et al.* 2005
- 20 research articles, 89 effects 'related to digital tools and learning environments to enhance literacy acquisition'. Weighted effect size of 0.49 indicating technology can have a positive impact on reading comprehension.

Klauer & Phye 2008
- 74 studies, 3,600 children. Training in inductive reasoning improves academic performance (0.69) more than intelligence test performance (0.52)

Gersten *et al.* 2009
- Maths interventions for low attainers. 42 studies ES ranging from 0.21-1.56. Teaching heuristics and explicit instruction particularly beneficial

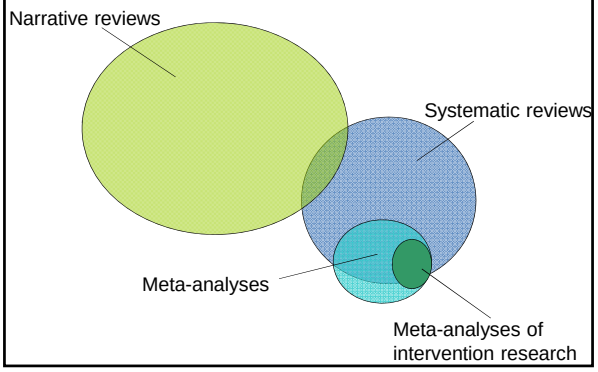
SRs in social science can address a range of RQs

- Impact or effectiveness questions (causal)
  - e.g., Does X work better than Y?
    - Homework intervention studies
- Correlational
  - e.g., examining strength of associations
    - Do schools with homework do better?
- Descriptive
  - e.g., describing and explaining participant perceptions and experiences
    - Describing teachers' and pupils' experiences and opinions about homework

Impact or effectiveness questions (causal)

- Intervention research
- Usually evaluation of policies, practices or programmes
- Usually based on experiments (randomised controlled trials or RCTs, quasi-experimental designs or QEDs)
- Answering impact questions
  - Does it work?
  - Is it better than…?

Literature reviews - conceptual relations



Which designs?

- RCTs only?
- RCTs plus *rigorously controlled* (high quality) experimental and quasi-experimental designs?
  - Individual RCTs
  - Cluster RCTs
  - Regression discontinuity designs
  - Quasi-experimental designs using a control population
  - Interrupted time series (IRS) designs
  - Prospective controlled cohort studies
- *All* RCTs and experimental designs?
- All pre-post comparisons?

Comparing a SR and a narrative review in the same topic

A systematic review of the research literature on the use of phonics in the teaching of reading and spelling

Torgerson, Brooks and Hall, 2006

Department for Education and Skills (DfES) commissioned the Universities of York and Sheffield to conduct a systematic review of experimental research on the use of phonics instruction in the teaching of reading and spelling. This review is based on evidence from randomised controlled trials (RCTs).

Early reading policy

Comparison of systematic review and narrative or 'expert' review

- Research question: Which is the most effective method for teaching children to read?
- In 2007 the UK Government's recommended method for teaching children to read at age 5 changed
- This followed the publication of two reviews in 2006 commissioned by the Department for Education, which tried to find out the answer to the research question using two different designs

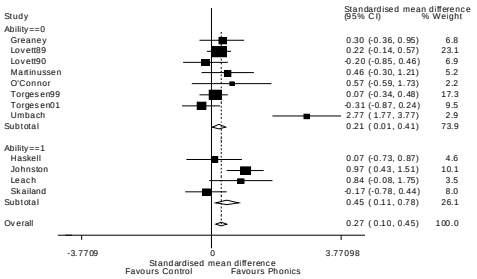
Two alternative designs

- Systematic review of different methods to teach reading
  - Used systematic review design
  - Included randomised trials undertaken anywhere
  - Examined the quality of the design of the included trials
  - Included a meta-analysis
  - Weighed up the evidence base before coming to conclusions.
- Narrative, expert review of different methods to teach reading
  - Used narrative, expert review design
  - Included examples of best practice, using a before and after design, one or two UK based trials, and expert opinion
  - Did not assess the quality of the design of the included studies
  - Conclusion derived from expert opinion.

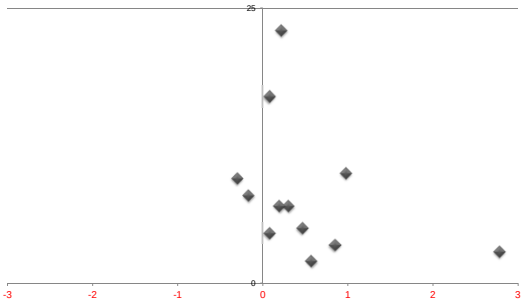
Strengths and limitations

- Systematic review design
  - Explicit, transparent, replicable methods
  - Systematic identification and inclusion of trials
  - Minimisation of bias at every stage in the design
  - Assessment of quality of evidence base
- Expert review design
  - No explicit methods, not replicable
  - Expert decision on study identification and inclusion
  - Potential for bias at every stage in the design
  - No assessment of the quality of the evidence base

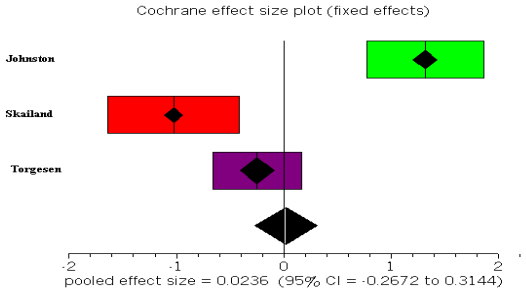
Meta-analysis: Forest plot



Publication bias: Funnel plot



Synthetic versus analytic phonics



Policy decision

"I am clear that synthetic phonics should be the first strategy in teaching all children to read."

Ruth Kelly, Secretary of State for Education and Skills

***Times* Mar. 21st 2006**

"The case for synthetic phonics is overwhelming."

Sir Jim Rose

***Times* Mar. 21st 2006**

"When the UK government recently introduced the 'synthetic phonics' method of teaching young children to read, they were told by Carole Torgerson, an evaluation expert at the University of York, that they could easily bolster the slim evidence base by randomising which schools joined the programme first."

***Financial Times* Mar.18th 2010**

"In 2007 the Government introduced a new reading strategy for primary schools based on synthetic phonics, which matches sounds to groups of letters. Professor Torgerson urged ministers to start a randomised trial: '…the introduction of phonics would be staggered, with schools chosen at random to start it one year or the next. Every child would have received the intervention, but it would have been possible to compare outcomes and establish whether phonics really works.' "

***Times* Sept. 24th 2011**

"Synthetic phonics does look promising," says Carole Torgerson of York University, one of the report's authors. "We found it had a moderate effect compared with whole-language approaches, but the evidence base for this conclusion was 12 relatively small trials, only one of which was UK-based. This would be an ideal time to do a national evaluation by implementing systematic synthetic phonics in some schools and not in others and then comparing the two."

***Economist* Mar. 26th 2006**

House of Commons Education and Skills Committee, 18th July 2005

"…in conducting his review, *Jim Rose will have the opportunity to draw on the findings of an independent systematic literature review of phonics use in the teaching and application of reading and spelling which we have commissioned from Professor Greg Brooks and Carole Torgerson*. This delivers on the public commitment we made…in 2003 to publish an analysis of existing research on phonics teaching methodologies. The aim … is to identify what is known from existing literature about how effective different approaches to phonics teaching are in comparison with each other, including the specific area of analytic versus synthetic phonics."

[Torgerson, Brooks and Hall, 2006]

Meta-analysis in a little more detail

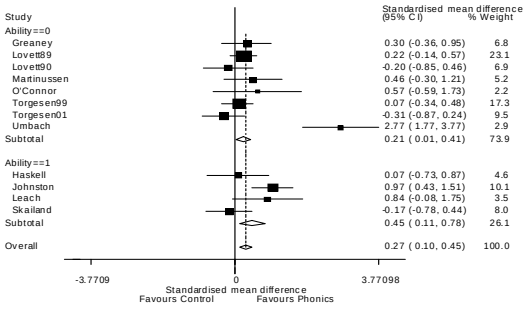
The rationale for using effect sizes

- Traditional reviews focus on statistical significance testing
  - Highly dependent on sample size
  - Null finding does not carry the same 'weight' as a significant finding
- Meta-analysis focuses on the ***direction*** and ***magnitude*** of the effects across studies
  - From '*Is there a difference?*' to '*How big is the difference?*'
  - Direction and magnitude represented by '***effect size***'

Forest plots

- Effective way of presenting results
  - Studies, effect sizes, confidence intervals
  - Provides an overview of consistency of effects
  - Summarises an overall effect (with confidence interval)
- Useful visual model of a meta-analysis

Forest plot



BREAK

One source of bias in SRs: Publication bias

What is publication bias?

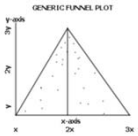
- The 'file drawer problem'
- Publication bias occurs when there are systematic differences in conclusions between studies that are unpublished compared with those that are published
  - statistically significant (positive) findings more likely to be published
  - smaller studies need larger effect size to reach significance
  - large studies tend to get smaller effect sizes
- Usually unpublished data are more likely to be 'negative' about an intervention than studies that are published.

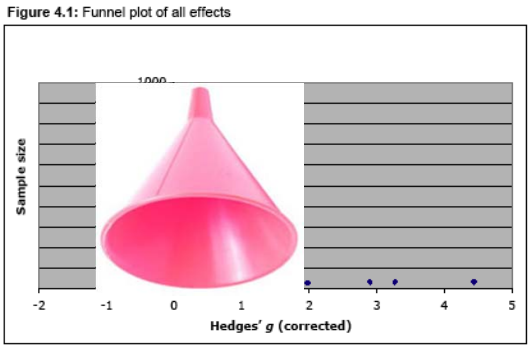
What are the effects of publication bias?

- In systematic reviews of randomised trials it is usual practice to put the trials into a meta-analysis (i.e., adding up all the studies)
- If only positive studies are published then we could erroneously conclude that an intervention was effective when, in truth, there was no benefit
- Replications difficult to get published

How can we detect publication bias?

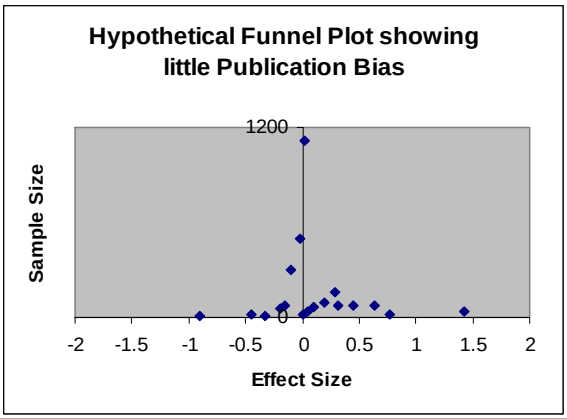
- One simple method of detecting its existence is through the use of a 'funnel plot'
- A funnel plot is a graphical device where all the effect sizes from individual studies are plotted on an x-axis whilst the size of the trial is plotted on the y-axis
- If there is NO publication bias the plots will form an 'inverted funnel'.



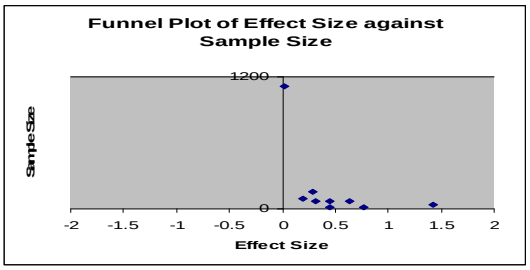


Why are negative studies not published?

- Researchers with negative studies may be disappointed in the results and not write them up and submit them for publication.
- Journal editors may refuse to publish negative studies.



Review of adult literacy teaching



Torgerson, Porthouse & Brooks. JRR, 2003

Critical appraisal of SRs

Not all SRs are equal: Critical appraisal of SRs

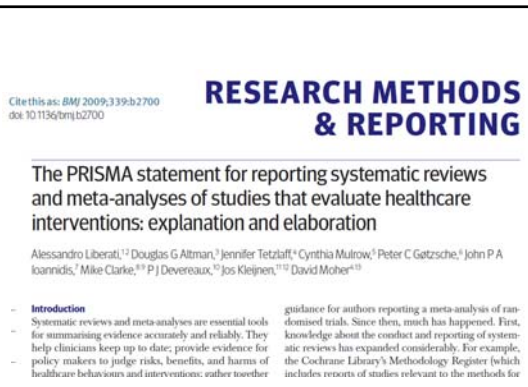
- PRISMA statement (**P**referred **R**eporting **I**tems for **S**ystematic Reviews and **M**eta-**A**nalyses)

http://www.prisma-statement.org/

- AMSTAR (**A**ssessment of **M**ultiple sy**ST**emat**A**tic **R**eviews)
- ROBIS

PRISMA-P

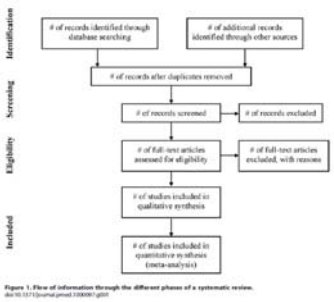
- Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P 2015)
  - Checklist to guide writing of SR protocols
  - 17 items considered essential and minimum components of a SR or meta-analysis protocol
  - Two key papers to read: the checklist and a paper providing an explanation of the items



PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	

Flow diagram



PRISMA-P papers

- Shamseer et al. Elaboration and explanation paper http://www.bmj.com/content/349/bmj.g7647
- Moher et al. Checklist paper http://www.systematicreviewsjournal.com/content/4/1/1/abstract

Methodological quality of meta-analyses on treatments for chronic obstructive pulmonary disease: a cross-sectional study using the AMSTAR (Assessing the Methodological Quality of Systematic Reviews) tool

Robin ST Ho[1], Xinyin Wu[1], Jinqiu Yuan[1], Siya Liu[1], Xin Lai[1], Samuel YS Wong[1] and Vincent CH Chung[1,2]

BACKGROUND: Meta-analysis (MA) of randomised trials is considered to be one of the best approaches for summarising high-quality evidence on the efficacy and safety of treatments. However, methodological flaws in MAs can reduce the validity of conclusions, subsequently impairing the quality of decision making.
AIMS: To assess the methodological quality of MAs on COPD treatments.
METHODS: A cross-sectional study on MAs of COPD trials. MAs published during 2000–2013 were sampled from the Cochrane Database of Systematic Reviews and Database of Abstracts of Reviews of Effect. Methodological quality was assessed using the validated AMSTAR (Assessing the Methodological Quality of Systematic Reviews) tool.
RESULTS: Seventy-nine MAs were sampled. Only 18% considered the scientific quality of primary studies when formulating conclusions and 49% used appropriate meta-analytic methods to combine findings. The problems were particularly acute among MAs on pharmacological treatments. In 48% of MAs the authors did not report conflict of interest. Fifty-eight percent reported harmful effects of treatment. Publication bias was not assessed in 65% of MAs, and only 10% had searched non-English databases.
CONCLUSIONS: The methodological quality of the included MAs was disappointing. Consideration of scientific quality when formulating conclusions should be made explicit. Future MAs should improve on reporting conflict of interest and harm, assessment of publication bias, prevention of language bias and use of appropriate meta-analytic methods.

npj Primary Care Respiratory Medicine (2015) **25**, 14102; doi:10.1038/npjpcrm.2014.102; published online 8 January 2015

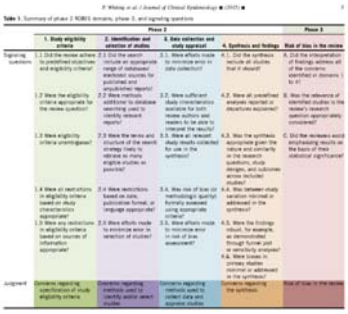
ROBIS

ELSEVIER

Journal of Clinical Epidemiology

ROBIS: A new tool to assess risk of bias in systematic reviews was developed

Penny Whiting[a,b,c,*], Jelena Savović[a,b], Julian P.T. Higgins[a,d], Deborah M. Caldwell[a], Barnaby C. Reeves[e], Beverley Shea[f], Philippa Davies[a,b], Jos Kleijnen[g,h], Rachel Churchill[a], the ROBIS group



Allocation bias in trials and its effects on systematic reviews

David Torgerson
Director, York Trials Unit
University of York

Allocation problems

- Allocation concealment is absolutely essential – some researchers/clinicians absolutely will subvert randomised allocation if possible – if this happens trial is damaged
- Large amounts of evidence in the health field that randomisation has been subverted (some evidence in criminal justice/social welfare as well)

Comparison good, poor randomisation

Allocation Concealment	Effect Size OR	
Adequate	1.0	
Unclear	0.67	P < 0.01
Inadequate	0.59	

Schulz et al. JAMA 1995;273:408.

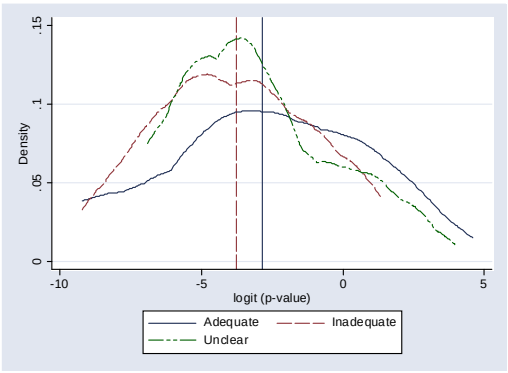
Mean ages of groups in a surgical trial

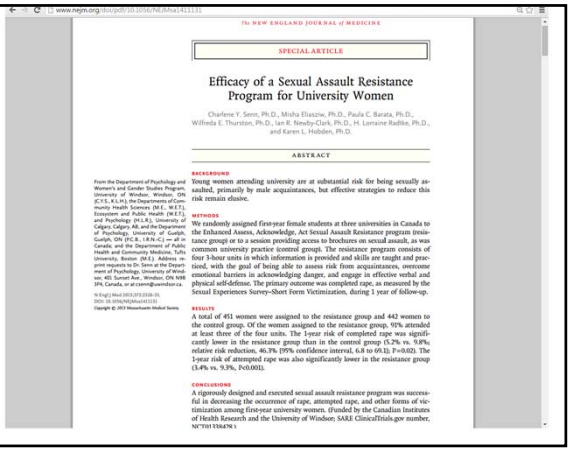
Clinician	Experimental	Control
All p < 0.01	59	63
1 p =.84	62	61
2 p = 0.60	43	52
3 p < 0.01	57	72
4 p < 0.001	33	69
5 p = 0.03	47	72
Others p = 0.99	64	59

More Evidence

- Hewitt and colleagues examined the association between p values and adequate concealment in 4 major medical journals
- Inadequate concealment largely used opaque envelopes
- The average p value for inadequately concealed trials was 0.022 compared with 0.052 for adequate trials (test for difference p = 0.045)

Hewitt et al. BMJ;2005: March 10th.





What is the problem here?

- There were 3 sites
- "Randomization was performed in permuted blocks of two with the use of the online tool Randomize.net, with stratification according to site"
- 452 assigned to control group and 464 to resistance group

Email correspondence

"Now it is me being confused. If you used a block of two stratified by site then the allocation will be perfectly balanced at each site every 2 women. If recruitment finished mid way through a block at each site then with 3 sites the biggest imbalance across the trial should be 3, shouldn't it?"
David

Dear David:
You are correct that, when the randomization process works perfectly, the maximum imbalance when stratified across 3 sites would be 3 subjects.

However, in practice, the computerized randomization process does not always work perfectly because of the human element. In our trial on several occasions, the research assistants mistakenly re-randomized subjects believing their online randomization had not been recorded or re-randomized subjects in an attempt to correct spelling mistakes, or mistakenly sent subjects to the wrong session.

Kindest regards

Last Email

Dear

You have a problem here. You need to inform the Journal of what you told me and write an addendum describing what happened. You have used an insecure system that the researchers could and did override, which can lead to bias. A systematic review would rate this allocation method as flawed.

Best wishes

David

NO RESPONSE

Does this affect systematic reviews?

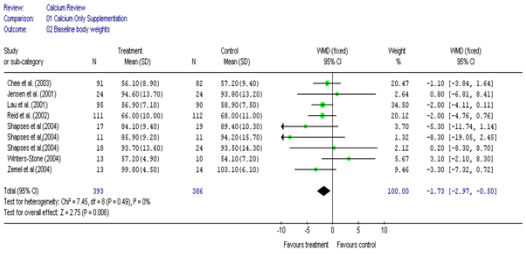
- If the problem of poor allocation practice were limited to a very few trials then, whilst there is a problem for some reviews, it shouldn't be a problem with majority of the evidence base
- Unfortunately, this may not be true

Systematic review of calcium for weight loss

- A systematic review of calcium supplements for weight loss – comparing body weights at final follow-up showed a statistically significant difference between the groups (-1.79 kg favouring calcium group; p = 0.005).
- But there was also a difference of baseline body weights.

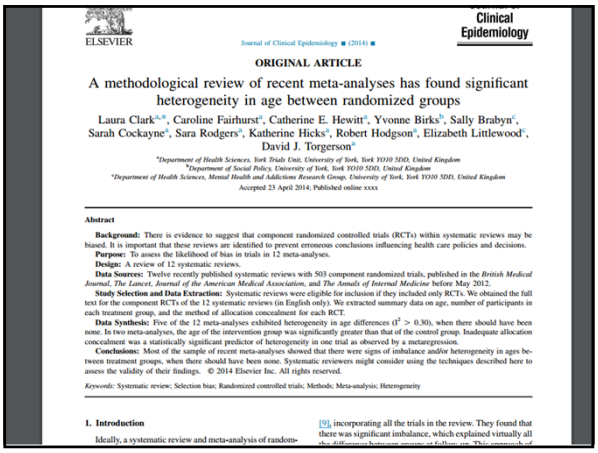
Trowman et al. Br J of Nutrition 2006;95:1033-38

Forest plot – baseline weight



Symptoms of bias

- Baseline variables should be balanced across trials. An individual trial might be in imbalance by chance but meta-analysis of several trials should generate an estimate close to zero with no heterogeneity
- If there is heterogeneity and or imbalance then some component trials could be biased and the whole review is tainted



ELSEVIER

Journal of Clinical Epidemiology ■ (2014) ■

Clinical Epidemiology

ORIGINAL ARTICLE

A methodological review of recent meta-analyses has found significant heterogeneity in age between randomized groups

Laura Clark[a,*], Caroline Fairhurst[a], Catherine E. Hewitt[a], Yvonne Birks[b], Sally Brabyn[c], Sarah Cockayne[a], Sara Rodgers[a], Katherine Hicks[a], Robert Hodgson[a], Elizabeth Littlewood[c], David J. Torgerson[a]

[a]Department of Health Sciences, York Trials Unit, University of York, York YO10 5DD, United Kingdom
[b]Department of Social Policy, University of York, York YO10 5DD, United Kingdom
[c]Department of Health Sciences, Mental Health and Addictions Research Group, University of York, York YO10 5DD, United Kingdom

Accepted 23 April 2014; Published online xxxx

Abstract

Background: There is evidence to suggest that component randomized controlled trials (RCTs) within systematic reviews may be biased. It is important that these reviews are identified to prevent erroneous conclusions influencing health care policies and decisions.
Purpose: To assess the likelihood of bias in trials in 12 meta-analyses.
Design: A review of 12 systematic reviews.
Data Sources: Twelve recently published systematic reviews with 503 component randomized trials, published in the *British Medical Journal*, *The Lancet*, *Journal of the American Medical Association*, and *The Annals of Internal Medicine* before May 2012.
Study Selection and Data Extraction: Systematic reviews were eligible for inclusion if they included only RCTs. We obtained the full text for the component RCTs of the 12 systematic reviews (in English only). We extracted summary data on age, number of participants in each treatment group, and the method of allocation concealment for each RCT.
Data Synthesis: Five of the 12 meta-analyses exhibited heterogeneity in age differences ($I^2 > 0.30$), when there should have been none. In two meta-analyses, the age of the intervention group was significantly greater than that of the control group. Inadequate allocation concealment was a statistically significant predictor of heterogeneity in one trial as observed by a metaregression.
Conclusions: Most of the sample of recent meta-analyses showed that there were signs of imbalance and/or heterogeneity in ages between treatment groups, when there should have been none. Systematic reviewers might consider using the techniques described here to assess the validity of their findings.

Keywords: Systematic review; Selection bias; Randomized controlled trials; Methods; Meta-analysis; Heterogeneity

1. Introduction

Ideally, a systematic review and meta-analysis of random-

[9], incorporating all the trials in the review. They found that there was significant imbalance, which explained virtually all

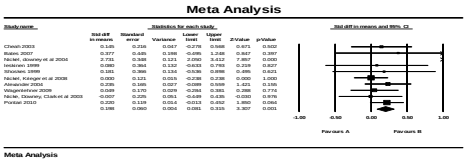
Why age?

- Two main reasons:
  - Easy characteristic for someone to use to subvert trial (e.g., older in control group)
  - Most trials will produce, by group, mean and SD of ages by allocated group

Review results ranked by I^2

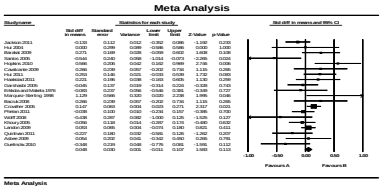
Systematic Review	Number of studies available for MA	Area	Intervention age mean (SD)	Control age. Mean (SD)	I squared value	P-value of difference in age
Anothaisintawee et al 2012	10	Drug	44.85 (5.56)	42.84 (5.67)	84.42	0.001
Rutjes et al 2012	38	Drug	62.17 (4.34)	62.44 (3.82)	67.92	0.835
Hemmingsen et al 2012	14	Drug	58.07 (4.13)	58.54 (3.98)	53.03	0.156
Thangaratinam et al 2012	20	Pregnancy and childbirth	28.15 (2.27)	27.95 (2.05)	50.11	0.113
Umpierre et al 2011	26	Lifestyle	58.29 (4.27)	58.79 (4.44)	42.72	0.173
Neumann et al 2012	9	Drug	64.18 (2.45)	63.94 (2.94)	33.46	0.029
Heneghan et al 2011	8	Other	63.15 (7.61)	62.71 (9.11)	31.62	0.024
Palmer et al 2012	11	Drug	51.99 (8.35)	52.86 (8.95)	29.03	0.173
Orrow et al 2012	10	Lifestyle	62.57 (10.29)	62.82 (9.72)	16.18	0.736
Coombes et al 2010	18	Drug	48.08 (6.9)	48.08 (7.25)	0.00	0.362
Leucht et al 2012	21	Drug	40.31 (9.24)	39.92 (9.78)	0.00	0.008
Hempel et al 2012	26	Drug	41.84 (24.43)	42.19 (25.24)	0.00	0.818

Heterogeneity: age difference



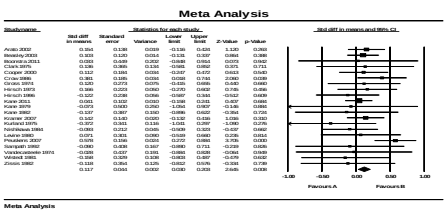
Difference in age p value = 0.001; I2 = 84.42

Heterogeneity: no age difference



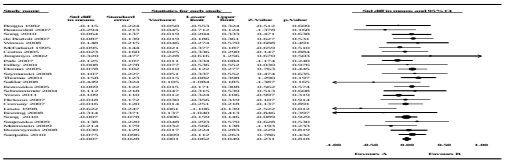
Difference in age p = 0.113; I2 = 50

Age imbalance no heterogeneity



Difference in age (p = 0.008;) hetererogeneity = 0.00

How it should be!



Difference in age p = 0.81; I2 = 0.00

Comment

- Out of 12 meta-analyses published in 4 leading medical journals:
  - Only 3 showed the expected zero heterogeneity and zero imbalance

A review conclusion

- In the review with >50% I2 it was concluded that:
  - *Dietary and lifestyle interventions can reduce maternal gestational weight gain and improve outcomes for both mother and baby*
- Is such a result believable – given the likelihood of biased trials?

Other predictors

- Age is the easiest but is not an outcome – what about covariates that more closely related to outcome (e.g., baseline pain scores of back pain)

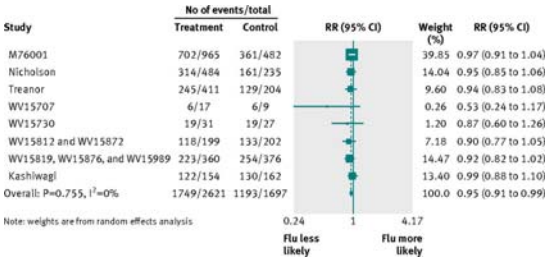
Hip fracture/backpain

Variable	Intervention group mean (SD)	Control group mean (SD)	I squared value	P-value for difference between groups	Meta regression
Age n=12	83.64 (6.91)	83.43 (7.12)	53.65	0.711	0.009
Body mass n=7	54.97 (11.47)	55.90 (11.39)	76.37	0.000	0.014

Variable	Intervention group mean (SD)	Control group mean (SD)	I squared value	P-value for difference between groups	Meta regression
Age n=20	44.17 (11.35)	43.98 (11.48)	21.99	0.911	0.191
Back pain n=17	17.11 (8.47)	17.18 (8.76)	55.79	0.505	0.395

Clark et al. J Clin Epidemiol 2015;68:175-81

Anti-body exposure to 'flu'



Study	No of events/total: Treatment	Control	RR (95% CI)	Weight (%)	RR (95% CI)
M76001	702/965	361/482		39.85	0.97 (0.91 to 1.04)
Nicholson	314/484	161/235		14.04	0.95 (0.85 to 1.06)
Treanor	245/411	129/204		9.60	0.94 (0.83 to 1.08)
WV15707	6/17	6/9		0.26	0.53 (0.24 to 1.17)
WV15730	19/31	19/27		1.20	0.87 (0.60 to 1.26)
WV15812 and WV15872	118/199	133/202		7.18	0.90 (0.77 to 1.05)
WV15819, WV15876, and WV15989	223/360	254/376		14.47	0.92 (0.82 to 1.02)
Kashiwagi	122/154	130/162		13.40	0.99 (0.88 to 1.10)
Overall: P=0.755, I^2=0%	1749/2621	1193/1697		100.0	0.95 (0.91 to 0.99)

Note: weights are from random effects analysis

0.24 1 4.17

Flu less likely | Flu more likely

Ebell 2013; Methodological concerns about studies on oseltamivir for flu (BMJ 2013;347:f7148)

Comment

- We have a BIG problem – the evidence suggests significant numbers of subverted trials are entering the 'food chain'
- Unless we scrap all the evidence of the last 50 years and start again what can we do?

Some suggestions/Discussion

- First routinely do baseline meta-analyses of age and another strong predictor of outcome SRs that pass this are likely to be OK
- Other suggestions:
  - Sort by baseline imbalance exclude those with a pre-specified baseline imbalance
  - Start a cumulative, by imbalance, meta-analysis stop when heterogeneity appears
  - Remove most severe studies in imbalance

References

- Chalmers, I., Hedges, L.V. and Cooper, H. (2002) 'A brief history of research synthesis', *Evaluation and the Health Professions*, *25*.
- Harris Cooper and Larry V. Hedges (Eds.) (1994) *The Handbook of Research Synthesis*. New York, N.Y: Russell Sage Foundation.
- Glass, G.V. (1976) 'Primary, secondary and meta-analysis', *Educational Researcher*, *5*.
- Glass, G.V., McGaw, B. and Smith, M.L (1981) *Meta-analysis in Social Research*. Beverly Hills, CA: Sage.
- Lipsey, M.W. and Wilson, D.B. (2001) *Practical Meta-analysis*. Applied Social Research Methods Series 49. London: Sage.
- Mulrow, C. (1994) 'Rationale for systematic reviews', *BMJ*, *309*.
- Pettricrew, M. (2001) 'Systematic reviews from astronomy to zoology: myths and misconceptions', *BMJ*, *322*.
- Torgerson, C. (2003) *Systematic Reviews*. London: Continuum.

Acknowledgements

- Some of the slides in this presentation are an outcome of the work of the ESRC-funded Researcher Development Initiative: *"Training in the Quantitative synthesis of Intervention Research Findings in Education and Social Sciences"* which ran from 2008-2011
- The training was designed by Steve Higgins, Rob Coe, Carole Torgerson (Durham University) and Mark Newman and James Thomas, Institute of Education, London University
- The team acknowledges the support of Mark Lipsey, David Wilson and Herb Marsh in preparation of some of the materials, particularly Lipsey and Wilson's (2001) "Practical Meta-analysis" and David Wilson's slides at: http://mason.gmu.edu/~dwilsonb/ma.html (accessed 9/3/11).