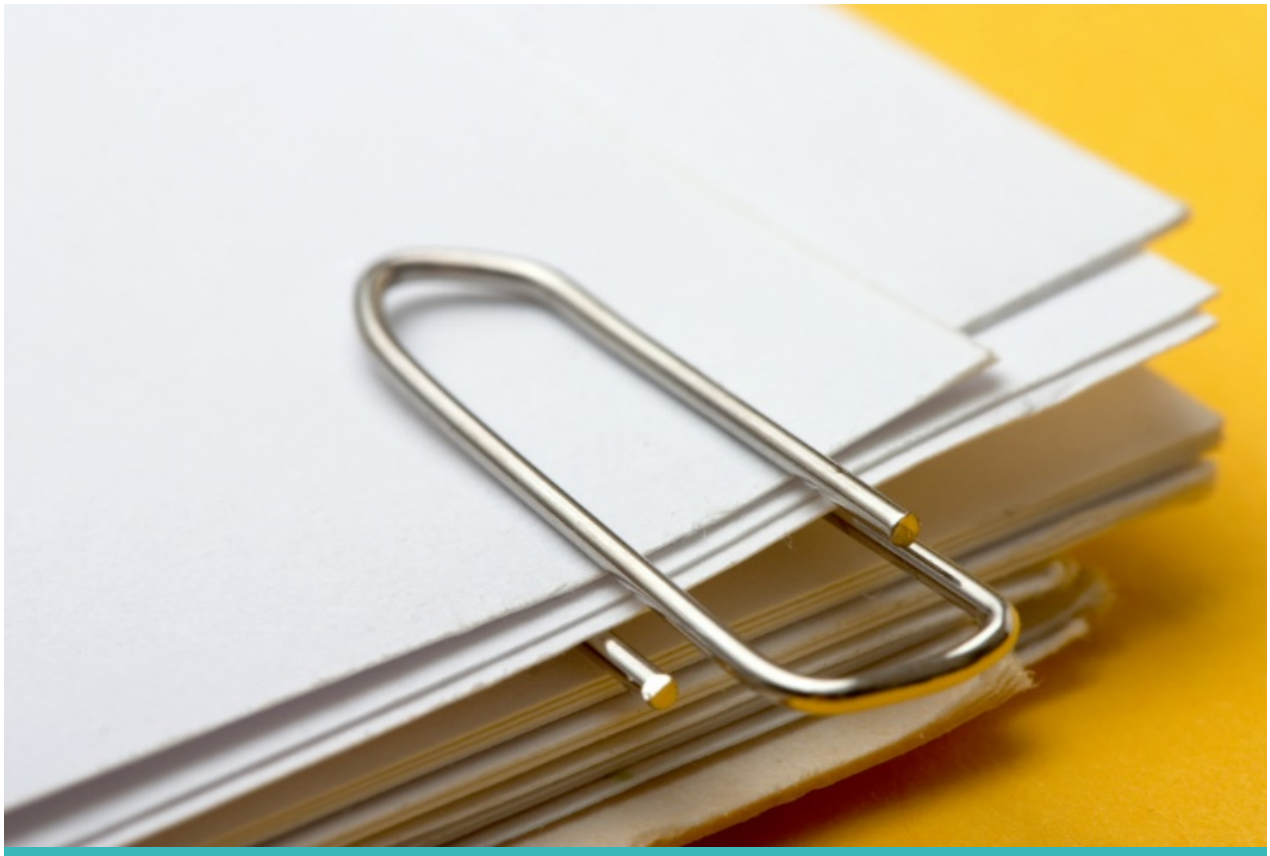
A guide to our
EVIDENCE REVIEW METHODS



Dawn Snape, Office for National Statistics
Catherine Meads, Brunel University London
Anne-Marie Bagnall, Leeds Beckett University
Olga Tregaskis, University of East Anglia
Louise Mansfield, Brunel University London
Sara MacLennan, DEFRA
Silvia Brunetti, What Works Centre for Wellbeing

Revised April 2019

Table of Contents

1. Purpose and approach to the Centre's evidence reviews

The What Works Centre for Wellbeing, in keeping with all members of the What Works Network, aims to produce high quality, accessible evidence syntheses for decision makers. Our evidence reviews will compare the effectiveness of different types of interventions or actions in improving wellbeing. To do this, we are developing a common currency for measuring wellbeing outcomes and will use a clear and consistent system for ranking the strength and quality of the evidence of what works to improve wellbeing as well as what doesn't work.

Each evidence synthesis will be designed for use by decision makers, and review questions will be developed and refined in consultation with the centre's users. Each review will provide, for each intervention, accessible and practical information about:

- Effectiveness and cost effectiveness of the intervention
- Applicability and implementation
- The strength of the evidence on which the assessment is based.

Due to insufficient evidence on some topics, it may not be possible to provide information on all of these aspects of each review. Where this is the case, we will identify research gaps and seek to work with partners to fill them. Each evidence programme will keep an evidence gap register for this purpose.

2. How 'wellbeing' is defined

The definition of wellbeing currently in use by the centre is based on the work of the Office for National Statistics:

> *Wellbeing, put simply, is about 'how we are doing' as individuals, communities and as a nation and how sustainable this is for the future. We define wellbeing as having 10 broad dimensions which have been shown to matter most to people in the UK as identified through a national debate. The dimensions are: the natural environment, personal wellbeing, our relationships, health, what we do, where we live, personal finance, the economy, education and skills and governance. Personal (or subjective) wellbeing is a particularly important dimension which we define as how satisfied we are with our lives, our sense that what we do in life is worthwhile, our day to day emotional experiences (happiness and anxiety) and wider mental wellbeing. (ONS, 2014)*

Different aspects of the definition and ONS wellbeing measurement framework will be relevant to different policy areas and services and to different evidence programmes. Across the centre as a whole, this definition and measurement framework will be an important starting point. For all evidence programmes, the definition and measures of personal wellbeing (also commonly referred to as subjective wellbeing) will form the basis of our approach to comparing evidence across different areas. All of our evidence reviews will look for evidence of how interventions and actions affect subjective wellbeing, no matter how it is measured. We will also look for evidence of wellbeing in other ways, including objective measures and measures that are relevant to specific topics (e.g. job satisfaction at work).

To enable comparisons of outcomes based on different measures of wellbeing, the centre will use life satisfaction as a common currency for subjective wellbeing. This does not mean that we will only look for direct evidence of effects on life satisfaction in our evidence reviews, but that it might ultimately be possible to convert evidence from other wellbeing measures into equivalent 'units' of life satisfaction to make it easier to compare wellbeing outcomes measured in different ways (a draft working paper with further information has been distributed to all evidence programmes and a revised version will be available shortly).

3. About our evidence reviews

The What Works Centre for Wellbeing conducts systematic and other forms of evidence reviews specifically to inform decision-making, with the aim of helping government, communities, business and people make better decisions to improve wellbeing.

The centre will take a variety of different approaches to reviewing evidence on wellbeing, depending on the nature and quality of evidence available. Usually, our work will entail systematic evidence reviews. According to the Cochrane Collaboration, 'a systematic review is a high-level overview of primary research on a particular research question that tries to identify, select, synthesize and appraise all high quality research evidence relevant to that question in order to answer it.' Important features of systematic reviews are that they:

- Collate all evidence that fits pre-specified eligibility criteria in order to address a specific research question; and
- Minimise bias by using explicit, systematic methods

Our systematic reviews may also incorporate meta-analysis where feasible and appropriate (see section 11 for further details). Meta-analysis is a statistical technique which combines the results of several different studies into a single numerical estimate of the effect size of an intervention.

4. Research methods appropriate to understanding wellbeing outcomes

The quality of evidence produced by a study will depend in large part on how appropriate the study design is for addressing the research question posed. The NICE public health guidelines provide a

helpful overview of different research designs best suited to answering different types of research questions.

The centre's work will focus on a range of issues requiring both quantitative and qualitative research designs, such as:

- measurement of the wellbeing impacts of different types of interventions or actions (interventional studies, particularly randomised controlled trials or any experimental study with a control group);
- the processes by which impacts occur (qualitative research);
- how and why impacts are experienced differently by different people or in different contexts (qualitative research);
- how interventions can be implemented most effectively (implementation studies);
- the cost-effectiveness of wellbeing interventions (economic studies).

Evidence standards and evidence quality checklists appropriate for assessing the quality of evidence from different research designs are therefore required. Details of the quality standards used by the centre can be found in sections 8 and 12 of this guide.

5. Planning evidence reviews and developing review protocols

The centre emphasises the importance of a clear, transparent and well-documented approach to each evidence review. The starting point is the development of a review protocol, documenting in advance the methods to be used in the review with the aim of minimizing bias and maximising transparency.

Recommended items to include in the review protocol are summarised in Box 1. For further details of what should be included in each section, see Systematic Reviews: CRD's guidance for undertaking reviews in health care, section 1.

All of the centre's systematic review protocols will be prospectively submitted to the Prospero database to register the details of the review. This will increase transparency and help to avoid possible duplication of work.

Box 1: Summary of what to include in a review protocol

Background: key contextual and conceptual factors relevant to the review question and the justification for the review.

The review question: state the main review question and any additional sub-questions to be addressed.

Study inclusion and exclusion criteria: clearly defined using PICOS/PECOS elements (see Box 2 for details).

Review methods to be used including:

Identification of research evidence
Selection of studies for inclusion
Data extraction for included studies
Quality assessment of included studies
Synthesis of results
Dissemination of the review findings

Process for making any protocol amendments: If any modification to the protocol is required after starting the work, protocol amendments should be clearly documented and justified. Details of how this will be done should be included in the original protocol.

(Adapted from Systematic Reviews: CRD's guidance for undertaking reviews in health care, p15)

6. Developing review questions

The nature and type of review questions determines the type of evidence reviews and the type of evidence that is most suitable (for example, intervention studies or qualitative data); both the type of evidence review and type of evidence need careful consideration (Petticrew and Roberts, 2003). The process for developing a review question is the same whatever the nature and type of question. Review questions should be clear and focused, with the exact structure of each question dependent on what is being asked.

The review question should specify the types of population (participants), types of interventions (and comparisons), and the types of outcomes that are of interest. The acronym PICOS (Participants, Interventions, Comparisons, Outcomes and Study Designs) helps to serve as a reminder of these. Box 1 highlights important questions to consider for each aspect of the PICOS framework. It is adapted from the NICE Review Guidelines which also provide further helpful information on developing review questions.

Box 2: Using the PICOS/PECOS framework to develop review questions

POPULATION: Which population are we interested in? How best can it be described? Are there subgroups that need to be considered? Where are the population? Which settings are they in?

INTERVENTION: Which intervention, treatment or approach should be used?

COMPARATORS: Are there alternative(s) to the intervention being considered? If so, what are these (for example, other interventions, standard active comparators, usual care or placebo)?

OUTCOME: Which outcomes should be considered to assess how well the intervention is working? What is really important for people using services?

STUDY DESIGNS: Which study designs should be included to address the research questions?

PECOS framework
Where the systematic review is not about an intervention but about **exposure**, for example to a risk factor, or an association between one factor and another, inclusion criteria of population, exposure, comparator, outcomes and study design should be used instead.

Specifying each of these aspects of the review question will form the basis of the pre-specified eligibility criteria for the review (Higgins & Green 2011). All of this information should be included in the review protocol. A more unusual example of a review question is systematic reviews of conceptual theories. More flexibility may be needed in defining inclusion criteria for these types of reviews.

In developing the review question, it is important to remember that people's wellbeing may be affected by interventions or changes in many areas of their lives. This suggests a need to keep inclusion criteria broad and to consider how a range of different study designs may provide relevant evidence.

It is important to consider the range of factors that may affect the outcomes and effectiveness of an intervention. This could include the design of the intervention itself, such as having multiple components addressing multiple causes and problems rather than taking a specific targeted approach, wider social factors that may affect wellbeing or any property of the system in which the intervention has been introduced. Thinking through the relevant sources of complexity at the beginning of the review will be helpful to identify important items that should be searched, extracted and synthesised. Logic models (also known as analytical frameworks) can be useful to map different components, causal pathways, as well as mediators and moderators of effect (Petticrew et al. 2018, Montgomery et al. 2019). In the (Montgomery et al. 2019) paper there are many examples of how taking a complexity perspective with the adoption of a logic model can help researchers to identify the different sources of heterogeneity and guide the extraction, analysis and interpretation of the evidence. In one of the

examples, the adoption of the complexity approach made it possible to identify and explain heterogeneity in the methodology and in the PICOS elements and decreased the risk for the evidence to be improperly downgraded.

In this context, it is also important to remember that the inclusion of complexity sources needs to be pragmatic and targeted as not all of the sources are relevant. Broad reviews can inevitably bring together many different interventions with different elements but it is recommended to focus on the active and key components which are relevant to the objectives of the review and the needs of the users.

For further details on adopting a complexity perspective for reviews of intervention effects, read section 13.1, Annex 6-7 and/or refer to the original paper (Montgomery et al., 2019).

Equity considerations should be included every part of the review protocol to ensure that wellbeing inequalities are captured throughout the work of the Centre. For equity-focused systematic reviews the way in which 'disadvantage' is defined should also be described if it is used as criterion in the review (e.g., for further information, see the PRISMA guidelines).

The setting for the question should also be specified if necessary. To help with this, outcomes and other factors that are important should be listed in the review protocol.

Box 3: Example of the use of PICOS criteria to specify the review question

Taken from a NICE mapping review on community engagement:

POPULATION: UK only. Communities involved in interventions to improve their health; health or social care practitioners or other individuals involved in developing, delivering or managing relevant interventions.

INTERVENTION: Focus on community engagement of any kind (for example, activities that ensure community representatives are involved in developing, delivering or managing services; or local activities that support community engagement). Local or national policy or practice.

COMPARISON: Studies with any or no comparators were eligible for inclusion.

OUTCOMES: improvement/ change in individual and population-level health and wellbeing; positive changes in health-related knowledge, attitudes and behaviour; improvement/ change in process outcomes (e.g. service acceptability, uptake, efficiency, productivity, partnership working); increase/ change in the number of people involved in community activities to improve health; increase in the community's control of health promotion activities; improvement in personal outcomes such as self-esteem and independence; improvement in the community's capacity to make changes and improvements to foster a sense of belonging; adverse or unintended outcomes; economic outcomes.

STUDY DESIGNS: Empirical research: either quantitative, qualitative or mixed methods outcome or process evaluations. To include grey literature and practice surveys. Published from 2000 onwards in English. Discussion articles or commentaries not presenting empirical or theoretical research will be excluded.

In some topic areas large numbers of systematic reviews already exist. In this case it may be more appropriate to conduct a systematic review of systematic reviews, rather than a systematic review of primary studies. A protocol will need to be written and inclusion criteria will still need to be defined, but the intervention criteria may need to be less tightly defined than with a systematic review of primary studies because there may need to be flexibility with interpretation. The risk of systematic review of systematic reviews is that some primary studies may be inadvertently double or triple counted whereas others may not be.

7. Searching for evidence

Search methods should aim to balance precision and sensitivity. The aim is to identify the best available evidence to address a particular question, without producing an unmanageable volume of results. This involves a forensic search that includes:

- creating precise search questions and identifying the study types needed to answer those questions
- considering synonyms of the search terms to enhance fuller retrieval of evidence

- matching key databases to the questions being asked (and not necessarily trawling all available databases just because they exist)
- adopting a pragmatic and flexible approach that allows a continual review of how best to find evidence
- having an understanding of the existing evidence base.
- using existing references that you already know about to make sure that you find them in your searches, demonstrating that your searches are adequate

All search processes should be transparent, clearly documented and reproducible. The search process itself should be as comprehensive as possible, bearing in mind time and resource limitations and should be based on a search protocol.

Search terms for wellbeing concepts are currently being developed and will include terms incorporating life satisfaction.

7.1 Developing a search protocol

The review team should develop a search protocol based on the review protocol. The search protocol sets out how evidence will be identified and provides a basis to develop a detailed search strategy. The search protocol is normally added as an appendix to each review protocol. Items to be included in the search protocol are shown in Box 4.

Box 4: What to include in the search protocol

Search question(s) and key concepts

Electronic sources to be searched (core, additional and economic databases plus any websites) and date ranges

Plans for additional searches (for example, citation or hand-searching)

Restrictions on searches (such as dates)

The centre will search globally for the best available evidence, but in keeping with practice among other What Works Centres, we will generally focus on studies conducted in countries with a similar level of GDP to the UK to maximise comparability. This restriction, and any exceptions to it, should be included in the search protocol.

Additionally, the centre will issue a call for evidence on the website prior to each review. This will extend the search as well as helping to build the evidence base by encouraging the centre's users to understand the types of evidence that are most helpful in understanding wellbeing. This should also be included in each search protocol.

7.2 Developing the search strategy

To develop a search strategy, each review team will 'translate' the concepts from the search protocol, including all the synonyms that will be used (thesaurus terms and free-text/keywords) into a plan specifying how they will search for evidence.

The search strategy needs to balance sensitivity (ability to identify relevant information) and specificity (precision – the ability to exclude irrelevant documents). However, the need for an exhaustive search (involving additional resources) also needs to be balanced against a more modest search that may miss some studies. The balance will depend on the nature of the review questions and the available evidence.

The review team then translates the search strategy (as necessary) for use with various databases. The results should be downloaded into reference management software. Items that cannot be downloaded into bibliographic software can be recorded in a Word document or spread sheet.

Searches should include a mix of: core databases, subject-specific databases and other resources, depending on the subject of the research question and the level of evidence sought. The databases searched must be relevant to the topic in terms of their coverage and content. Where there are a large number of possibilities, it would be expedient to prioritise those most likely to produce relevant evidence. (For example, MEDLINE is unlikely to be a useful source of information for a review of social and emotional wellbeing in primary education, but ERIC would be.)

Study-type limits or filters should not be used, due to the broad nature of wellbeing evidence and the fact that the majority of sociological and social science databases do not provide adequate indexing by study design, and the quality of indexing for – and the vocabulary used in – study methodologies and designs varies extensively and, in some instances, is poor.

The start date for searches is determined by the nature of the evidence base and the time available to process data and the rationale should be documented in the search protocol.

For further details on developing a search strategy for systematic reviews, read section 6.4 of the Cochrane Handbook for systematic reviews of interventions (Lefebvre et al. 2011).

7.3 Conducting searches in topic areas relevant to wellbeing

7.3.1 Public health related searches

Searching for evidence on public health related topics may be long and complex and can present a technical challenge due to the nature of the databases available. Public health information resources do not use a standard indexing vocabulary or thesaurus and the thesauruses used by clinical databases only cover a limited number of public health concepts. The use of natural language varies, and studies, outcomes, measures and populations are not described in a consistent way.

The broad multidisciplinary nature of public health means that searches are carried out across a wide range of databases – currently, there are no dedicated national databases that bring this information together.

Websites can be a useful source of grey literature for public health reviews, particularly as a search of traditional, peer-reviewed literature may not produce much information. Careful selection of websites is

required to ensure that the type of evidence available is likely to be relevant: finding relevant data is more important than doing an exhaustive search.

As there may be a lack of particular types of evidence, such as controlled trials, this may limit the methodological coverage of systematic reviews if the review process follows the most rigorous evidence-based standards. There needs to be a balance so that the best evidence that is available can be included. This entails using a hierarchy so that, for effectiveness of interventions for example, if there is no randomised controlled trial evidence, cohort study evidence is used, and if no cohort evidence then case-control study evidence is used etc.

7.3.2 Economic searches

It is advisable to develop a fairly simple search strategy for economic searches because a complex search may exclude relevant studies. For example, instead of searching for population group and setting and intervention and the problem, it might be more reliable to just search for the public health problem. If this produces too many results, then additional concepts can be added.

Economic evidence searches can be undertaken using several existing databases. Examples include the NHS Economic Evaluation Database (EED) which accessible via the Cochrane website, EconLit, and Research Papers in Economics (RePec) The latter also includes a contact alert for new economics papers on happiness. MEDLINE also has some economics papers. Economic evidence can also be identified when sifting effectiveness or qualitative search results.

7.4 Extending the search

If the main searches have not retrieved all of the relevant material, the review team may need to widen the search and carry out additional types of searches. These could include: 'snowballing' to find citations, a search of the grey literature, journal hand-searches or making contact with experts and stakeholders.

7.4.1 Citations using 'snowballing'

A search can be usefully extended by looking for articles that cite other, more specific articles containing additional relevant references. However, it depends on whether the database software can perform this search; even if it is possible, such a search will only retrieve cited articles from journals indexed in the same database.

7.4.2 Grey literature

Grey literature is research that has not been published in a peer reviewed journal. Often it is research in the form of reports on the internet, but usually does not have an ISSN or ISBN number and is often not indexed in the searchable research databases such as Medline. It is important to consider grey literature alongside literature which has been published in a peer reviewed journal, because a search of the 'grey literature' can help identify material that will not be picked up by mainstream sources (such as the MEDLINE database). There is a tendency for studies with positive findings to be published earlier than those with negative or equivocal findings and be more likely to be published at all in peer-reviewed journals - publication bias. If we only look for journal articles, the publication bias of mostly finding positive studies is magnified and perpetuated in our systematic reviews. Grey literature is not constrained by research publication requirements and thus offers a wide range of works at varying stages of development which can provide further insight for a potentially more diverse range of users.

Grey literature should be subject to the same inclusion criteria as academic literature. It is important to consider the state of the evidence, the potential benefits from including grey literature, while being proportionate. Where grey literature is not included in the systematic review, it is important to clearly state the reasons.

Grey literature databases include OpenSIGLE and OAISTER. Both a database and an Internet search (on Google, for example) may be necessary, and calls for evidence will be issued via the Centre's website, but it is essential to be clear about the type of material needed. In particular, it is useful to distinguish between data that might supplement the effectiveness literature (for example, ongoing evaluative research) and information that could aid implementation. Grey literature should only be included in a review if the source can be cited i.e. details of the authors (whether individuals or institution/group), and publisher are given.

7.4.3 Hand-searching

Hand-searching involves a manual search through the contents tables of selected journal titles for relevant articles. There is no requirement to do this and it can be time consuming. However, it is worth doing if the reviewers are aware of any relevant journal titles that are not included in the bibliographic databases being searched. Hand-searching can also be worthwhile if the database searches have failed to retrieve much relevant evidence (though it should be limited to a few relevant, specialised journals). Bibliographic details of any studies identified should be added manually to the database of references that have been downloaded.

7.4.4 Contacting experts

Some types of research, notably intervention trials, are often documented in databases of ongoing research. However, these are not always up-to-date and it is advisable to ask experts in the area. Experts can be identified and contacted via research networks, relevant journal abstracts or via relevant reference lists. Any additional evidence received should be entered into the bibliographic database. The number of articles identified by this means must be specified in the methods section of the review.

7.4.5 Using review-level material to identify primary studies

Review-level material (for example, systematic reviews, literature reviews and meta-analyses) may provide an additional source of primary studies. Relevant reviews can be identified using an appropriate checklist. The reference lists in the reviews can be used to identify potentially relevant primary studies.

The Centre for Reviews and Dissemination (CRD), Cochrane and Campbell databases are useful sources of robust, quality reviews.

7.4.6 What to do if your searches find little or no relevant evidence

A systematic review is intended to answer an important question around wellbeing. If there is little or no evidence on that important question, this is useful information that needs to be disseminated as it indicates that more research is needed in this area. These gaps in the evidence base can be collated as a research gap register which can then be used to plan future research programmes.

If the inclusion criteria are relaxed slightly it may be that more evidence can be found, but it tends to be of lower quality or doesn't quite answer the question raised. For example you may decide that there were no comparative studies, in which case single group studies are the only relevant evidence, even though they may be of very little help in determining whether an intervention is effective because of confounding factors. In a systematic review of an intervention for children you may find that there is little or no evidence in children but some in young people under the age of 25. In your systematic review you may be interested in a specific sort of subjective wellbeing outcome, but find that none of your studies

measured this, but did measure other outcomes such as depression or attendance. In this case it would be useful to report these instead and be explicit about the lack of wellbeing outcomes.

7.5 Documenting the search process

Systematic literature searches should be thorough, transparent and reproducible to minimise 'dissemination biases' (Song et al 2010). For these reasons, as well as to aid quality assurance, it is important to document it. The review team should be able to provide the following, once the searches are complete:

- Word document containing the search strategies for each resource searched.
- Final de-duplicated Endnote (or other reference management software) database of “hits”
- Word document of other results (for those records that cannot be downloaded into EndNote such as website results).

Box 5 summarises a best practice approach to searching for evidence and documenting the search, based on the PRISMA guidance (Welch et al, 2015).

Box 5: Documenting the search for evidence

For all evidence searches:

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.

Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.

Additionally, for systematic reviews including equity-related questions:

Describe the broad search strategy and terms used to address equity questions of the review.

Describe information sources (e.g., health, non-health, and grey literature sources) that were searched that are of specific relevance to the equity questions of the review.

8. Selecting studies for inclusion in the reviews

This section applies to both qualitative and quantitative evidence reviews and is based on the NICE public health systematic review guidance and PRISMA guidance.

Identifying and selecting all relevant studies is a critical stage in the evidence review process. Before undertaking screening, the review team should discuss and work through examples of studies meeting the inclusion criteria (as set out in the agreed review protocol) to ensure a high degree of inter-rater

reliability. Then studies meeting the inclusion criteria should be selected using the 2-stage screening approach below:

> **Stage 1: Title or abstract screening.** Titles or abstracts should normally be screened independently by 2 reviewers (that is, they should be double-screened) using the parameters set out in the review protocol. If the number of titles and abstracts retrieved is very large, a random selection (eg, 20%) may be double-screened, with the remainder being single screened. Any disagreements or queries about a study's relevance should be resolved by discussion with the other reviewers. If, after discussion, there is still doubt about whether or not the study meets the inclusion criteria, it should be retained.
>
> **Stage 2: Full-paper screening**: once title or abstract screening is complete, the review team should assess full-paper copies of the selected studies, using a full-paper screening tool developed for this purpose. This should normally be done independently by 2 people (that is, the studies should be double-screened). Any differences should be resolved by discussion between the 2 reviewers or by recourse to a third reviewer.

The study selection process should be clearly documented and include details of the inclusion criteria.

For example, this should 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 the rationale. In addition, for equity-focused systematic reviews, describe the rationale for including particular study designs related to the equity research questions.

A flow chart should be used to summarise the number of papers included and excluded at each stage of the process and this should be presented in the review report. The PRISMA flow diagram is a good example (also available in Annex 1).

Each study excluded at the full-paper screening stage should be listed in the appendix of the review, along with the reason for its exclusion.

9. Software to help with systematic reviews

There is a variety of software that can help with systematic reviews. Commonly used programmes are:

- Endnote, Reference Manager, RefWorks and other bibliographic software. This type of software can be used to download the searches, sift through studies and keep track of inclusion decisions etc.
- Systematic reviewing software such as RevMan (freely downloadable from the Cochrane Library website) and EPPI software. This can be useful for more of the systematic reviewing procedures than searches and reference management and is often used for data extraction. Data extraction can also be done in Excel or other spreadsheet packages.
- Meta-analysis software or packages that can do meta-analyses, such as STATA and Comprehensive Meta-analysis. NB Revman can also do very good meta-analyses.

10. Data extraction

Data extraction of each full paper into a pre-agreed form or evidence table should be undertaken by one reviewer and checked for accuracy by another. Periodically throughout the process of data extraction, a random selection should be considered independently by 2 people (that is, double-assessed). The size of the sample will vary from review to review, but a minimum of 10% of the studies should be double-assessed. Any differences should be resolved by discussion or recourse to a third reviewer.

For all reviews, the evidence table should list and define all variables for which data were sought (e.g., PICOS, numerical results, funding sources) and any assumptions and simplifications made.

Where given, exact p values (whether or not significant) and confidence intervals must be reported, as should the test from which they were obtained. For the centre's evidence reviews, a p value of ≤0.05 is considered statistically significant. Where p values are inadequately reported or not given, this should be stated. Any descriptive statistics (including any mean values) indicating the direction of the difference between intervention and comparator should be presented. If no further statistical information is available, this should be clearly stated. Where study details are inadequately reported, absent (or not applicable), this should also be clearly stated.

In addition, for equity-focused systematic reviews, all data items related to equity should be listed and defined (e.g., using PROGRESS-Plus or other criteria, context). For further details, see Welch et al, 2015.

Box 6 lists the key items that should be included in the evidence table.

Box 6: Information to include in an evidence table

Bibliography (authors, date)
Study aim and type (for example, RCT, case–control)
Population (source, eligible and selected)
Intervention, if applicable (content, intervener, duration, method, mode or timing of delivery)
Method of allocation to study group (if applicable)
Numbers of participants in each group at baseline and at follow up (if applicable)
Outcomes (primary and secondary and whether measures were objective, subjective or otherwise validated)
Key numerical results (including proportions experiencing relevant outcomes in each group, means and medians, standard deviations, ranges and effects sizes)
Inadequately reported or missing data.

10.1 Foreign Language papers

Even where searches include foreign languages, usually less than 1% of potentially includable papers are in written entirely in foreign languages. Where you have a paper in a foreign language you may frequently have an abstract in English which can be used to decide whether it is includable according to your inclusion criteria. If you consider that it is includable we don't recommend that you have the paper formally translated. This is because you frequently don't need the whole paper, just the methods and results, translators often don't know the technical language so you may need to ask further questions to understand the translation, often the table and figure legends don't get translated, and it is expensive. Instead we suggest that you find someone who speaks that language and meet with them. Ask them to read the paper in advance then ask them specific questions in order to complete your data extraction and quality assessment sheets. That way you can explain to them the technical issues you are looking for and they can describe much more clearly what is actually on the paper.

11. Assessing the quality of the evidence

The review team should assess the quality of evidence selected for inclusion in the review using the appropriate quality appraisal checklist. Quality assessment is a critical stage of the evidence review process.

Before undertaking the assessment, the review team should discuss and work through some of the studies to ensure there is a high degree of inter-rater reliability. Each full paper should be assessed by one reviewer and checked for accuracy by another.
Periodically throughout the process, a random selection should be considered independently by 2 people (that is, double-assessed). The size of the sample will vary from review to review, but a minimum of 10% of the studies should be double-assessed. Any differences in quality grading should be resolved by discussion or recourse to a third reviewer.

Some studies, particularly those using mixed methods, may report quantitative, qualitative and economic outcomes. In such cases, each aspect of the study should be separately assessed using the appropriate checklist. Similarly, a study may assess the effectiveness of an intervention using different outcome measures, some of which will be more reliable than others (for example, self-reported anxiety versus a measure of cortisol levels in blood samples). In such cases, the study might be rated differently for each outcome, depending on the reliability of the measures used. For further information on how to integrate evidence from qualitative and quantitative studies, see Dixon-Woods et al (2004).

External validity (also known as generalisability) is how well the evidence in the research you are assessing can be relevant to the situation locally in the UK. Some research may not be locally relevant because, for example, the setting is completely different, or the intervention might not be locally acceptable. This is very much a matter of judgement and if it doubt, you may need to come to a consensus within the team.

11.1 Checklists to use for assessing evidence quality

The Centre will use specific evidence quality checklists for qualitative and quantitative research designs. The quality of evidence from each primary study (and different aspects of the same study in

the case of mixed methods designs) should be assessed using the relevant quality checklist (see Box 7).

Each individual aspect of the study is given a quality rating based on the criteria included in the checklist.

For qualitative research, an assessment must be made of the methodological strengths and weaknesses of each study as there is no hierarchy of study design within qualitative research. Review authors should present and explain these assessments in documenting the review process.

Box 7: Quality checklists for different types of evidence

For **quantitative evidence** of intervention effectiveness, use the checklist of evidence quality adapted from the Early Intervention Foundation in Annex 2.

For **qualitative evidence**, use the checklist adapted from CASP in Annex 3.

11.2 Reporting evidence quality

Each individual aspect of the study is given a quality rating based on the criteria included in the checklist. This should be used to comment briefly on the risk of bias of each study, within the study summary table. This column will help to inform the final GRADE / CerQUAL judgement of the body of evidence, along with the other factors (discussed below).
Each study should not be given an overall 'quality rating'.

12. Evidence synthesis and meta-analysis

Both qualitative and quantitative evidence reviews should incorporate narrative summaries of, and evidence tables for, all studies. Concise detail should be given (where appropriate) on:

- population and settings
- interventions and comparators
- outcomes (measures and effects).

This includes identifying any similarities and differences between studies, for example, in terms of the study population and setting, interventions, comparators and outcome measures.

Results from relevant studies (whether statistically significant or not) can be presented graphically. It may also be useful to relate the evidence to logic models or theories of change.

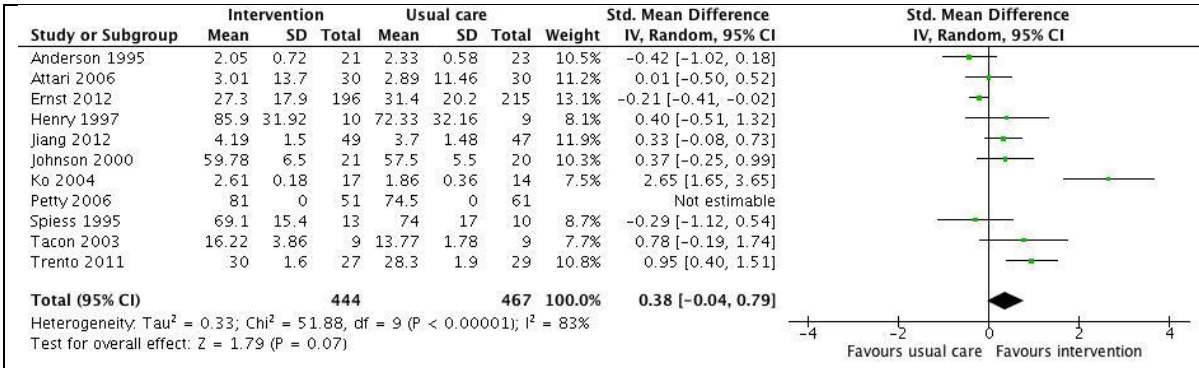
12.1 Using meta-analysis and other graphical methods of reporting

Meta-analysis is the pooling of numerical outcome results from different studies together into one plot and deriving an overall numerical estimate of effect size. It is usually presented in a Forest plot. (see Box 8 for an example). When considering doing meta-analysis, it's advisable to consult an expert.

12.1.1 Deciding when to use meta-analysis

Meta-analysis is appropriate when the same entity is being measured in similar populations in different studies and the comparators are also similar. For example, if you have a collection of 3 or more controlled trials where a similar intervention has been implemented, the controls are similar and the outcome measure, such as anxiety has been reported. Anxiety can be measured in a variety of ways, such as different questionnaire measures of anxiety, interview scales etc. It can also be reported in a variety of way – categorical (percentage above or below a specific cut-off point in the scale) or continuous (mean and standard deviation). In a single Forest plot, categorical and continuous measures cannot be combined.

Box 8: Example of a Forest Plot



Study or Subgroup	Intervention Mean	SD	Total	Usual care Mean	SD	Total	Weight	Std. Mean Difference IV, Random, 95% CI
Anderson 1995	2.05	0.72	21	2.33	0.58	23	10.5%	-0.42 [-1.02, 0.18]
Attari 2006	3.01	13.7	30	2.89	11.46	30	11.2%	0.01 [-0.50, 0.52]
Ernst 2012	27.3	17.9	196	31.4	20.2	215	13.1%	-0.21 [-0.41, -0.02]
Henry 1997	85.9	31.92	10	72.33	32.16	9	8.1%	0.40 [-0.51, 1.32]
Jiang 2012	4.19	1.5	49	3.7	1.48	47	11.9%	0.33 [-0.08, 0.73]
Johnson 2000	59.78	6.5	21	57.5	5.5	20	10.3%	0.37 [-0.25, 0.99]
Ko 2004	2.61	0.18	17	1.86	0.36	14	7.5%	2.65 [1.65, 3.65]
Petty 2006	81	0	51	74.5	0	61		Not estimable
Spiess 1995	69.1	15.4	13	74	17	10	8.7%	-0.29 [-1.12, 0.54]
Tacon 2003	16.22	3.86	9	13.77	1.78	9	7.7%	0.78 [-0.19, 1.74]
Trento 2011	30	1.6	27	28.3	1.9	29	10.8%	0.95 [0.40, 1.51]
Total (95% CI)			**444**			**467**	**100.0%**	**0.38 [-0.04, 0.79]**

Heterogeneity: Tau² = 0.33; Chi² = 51.88, df = 9 (P < 0.00001); I² = 83%
Test for overall effect: Z = 1.79 (P = 0.07)

This is a Forest plot of a continuous measure where each study has measured the outcome in a different way so standardised mean difference has been used. The vertical line in the plot is the line of no significant difference. The outcome for Petty 2006 is not estimable because no standard deviations were reported. The combined meta-analysis result is shown in the diamond and is SMD 0.38 (95% confidence intervals -0.04 to + 0.79. As this crosses the vertical line, it shows no significant difference.

Meta-analysis is ***not*** appropriate when:

- the populations are very different (eg, two studies are in adults and one is in children and the effects in children are very different)
- the interventions are different
- the comparators are different (such as no intervention in one study, an active intervention which is known to have a beneficial effect in another)
- the outcomes measured are different (eg, depression in one study, negative affect in another)

If meta-analysis is not appropriate in your study, there are other ways of graphically presenting your results such as a Harvest plot [Ogilvie et al. 2008]). Another alternative is to present the Forest plot without a combined estimate of effect size (ie, omit the bottom line with the diamond in the plot).

12.1.2 Heterogeneity and meta-analysis

The variability between studies is called heterogeneity and can refer to differences between populations, settings, interventions, comparators, outcomes and study designs. When these vary between study this is known as clinical heterogeneity, as opposed to statistical heterogeneity which is the statistical variation between studies. For example the Forest plot in the example is showing statistical heterogeneity in that some of the effect size estimates for the individual studies (the horizontal lines) vary in where they are on the plot, some are crossing the vertical line and Ko 2004 is very much towards the RHS. This statistical heterogeneity could be driven by clinical heterogeneity. Ko 2004 is from China so the population in that trial may be very different from those in the other trials.

Statistical heterogeneity is measured in meta-analysis by the Chi^2 test and by the I^2 test. In the example above the Chi^2 test was 51.88 for 9 degrees of freedom (df is the number of studies -1). The p value for the Chi^2 test was much less than 0.05 so there was significant statistical heterogeneity. The I^2 test can vary from 0% to 100% where 0% is no heterogeneity and 100% is maximum heterogeneity. In this example it was 83% which is considerable statistical heterogeneity. For methodological heterogeneity (for example, where trials of varying quality are involved), sensitivity analyses can be carried out by varying the number of studies in the meta-analysis.

Where there is a considerable amount of heterogeneity, meta-analysis can be conducted using a random effects model, which accounts for heterogeneity to some extent. Alternatively the impact of known research heterogeneity (for example, population characteristics or the intensity or frequency of an intervention) can be managed using methods such as subgroup analyses and meta-regression. Considerable heterogeneity can be a reason for not conducting meta-analysis at all. This is a matter that is under academic dispute somewhat so please refer to an expert in meta-analysis if you are unsure about how to deal with statistical heterogeneity in your meta-analysis. In the example above, a random effects model was used and the meta-analysis regarded as exploratory.

12.1.3 Dealing with missing data

Forest plots should include lines for studies that are believed to contain relevant data, even if details are missing from the published study. An estimate of the proportion of missing eligible data is needed for each analysis (as some studies will not include all relevant outcomes).

Sensitivity analysis can be used to investigate the impact of missing data. When outcome measures vary between studies, it may be appropriate to present separate summary graphs for each outcome. However, if outcomes can be transformed on to a common scale by making further assumptions, an integrated (graphical) summary may be helpful. In such cases, the basis (and assumptions) used should be clearly stated and the results obtained in this way should be clearly indicated.

12.2 Reporting the results of evidence synthesis and meta-analysis

The characteristics and limitations of the data in a meta-analysis should be fully reported (for example, in relation to the population and setting, intervention, sample size and validity of the evidence).

The methods of handling data and combining results of studies, if done, including measures of consistency for each meta-analysis should also be described.

In addition, for equity-focused systematic reviews, the methods of synthesizing findings on inequities (e.g., presenting both relative and absolute differences between groups) should also be described.

12.2.1 Assessing possible sources of bias

Publication bias (studies, particularly small studies, are more likely to be published if they include statistically significant or interesting results) should be critically assessed and reported. It may be helpful to inspect funnel plots for asymmetry to identify any publication bias (see the Cochrane website; also Sutton et al 2000).

Similarly, the possibility of selective reporting of outcomes (emphasising statistically significant results over others, for example) should be considered. In part, this can be done by examining which outcomes were described as primary and secondary in study reports or protocols.

A full description of data synthesis, including meta-analysis and extraction methods, is available in: Undertaking systematic reviews of research on effectiveness (NHS Centre for Reviews and Dissemination 2009).

12.2.2 Assessing applicability

The review team should use the quality appraisal checklist to assess the external validity of quantitative studies: the extent to which the findings for the study participants are generalizable to the whole 'source population' that they were chosen from. This involves assessing the extent to which study participants are representative of the source population. It may also involve an assessment of the extent to which, if the study were replicated in a different setting but with similar population parameters, the results would have been the same or similar. If the study includes an 'intervention', then it will also be assessed to see if it would be feasible in settings other than the one initially investigated. Most qualitative studies by their very nature will not be generalizable. However, where there is reason to suppose the results would have broader applicability they should be assessed for external validity.

The following characteristics should be considered:

POPULATION: Age, sex/gender, race/ethnicity, disability, sexual orientation/gender identity, religion/beliefs, socioeconomic status, health status (for example, severity of illness/ disease), other characteristics specific to the topic area/review question(s).

SETTING: Country, geographical context (for example, urban/rural), legislative, policy, cultural, socioeconomic and fiscal context, other characteristics specific to the topic area/review question(s).

INTERVENTION: Feasibility (for example, in terms of available services/costs/reach), practicalities (for example, experience/training required), acceptability (for example, number of visits/ adherence required), accessibility (for example, transport/outreach required), other characteristics specific to the topic area/review question(s).

OUTCOMES: Appropriate/relevant, follow-up periods, important effects on wellbeing. You may also need to report wellbeing results by protected characteristic group (ie subgroup analyses by protected characteristic) if available.

13. Rating the quality of the evidence for each finding in a review

To help decision-makers understand the degree of confidence they can have in the findings from the Centre's evidence reviews, a rating will be provided of the overall quality of the evidence for each individual finding in the reviews.

The GRADE and CERQual approaches will be used to assess and rate the quality of evidence for specific findings in both quantitative and qualitative evidence reviews, respectively. The GRADE and CERQual methodologies are well-documented, in widespread use, and provide clear approaches to rating the quality of evidence for findings within a review. They also provide a transparent approach to rating the strength of any recommendations made on the basis of the review findings.

13.1 Use of GRADE to rate the quality of evidence for findings in quantitative reviews

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach will be used for grading the quality of evidence from quantitative systematic reviews. It has been adopted by over 60 organisations internationally including the Cochrane Collaboration and other What Works Centres (eg, NICE) and is increasingly recognised as best practice. The Cochrane Handbook describes the approach in the following way:

> *"…the GRADE approach defines the quality of a body of evidence as the extent to which one can be confident that an estimate of effect or association is close to the quantity of specific interest. Quality of a body of evidence involves consideration of within-study risk of bias (methodological quality), directness of evidence, heterogeneity, precision of effect estimates and risk of publication bias…The GRADE system entails an assessment of the quality of a body of evidence for each individual outcome." (*Cochrane Handbook).

There are four quality level ratings used in the GRADE approach as shown in Box 9.

Box 9: How quality is defined using the GRADE approach

HIGH QUALITY: Further research is very unlikely to change our confidence in the estimate of effect

MODERATE QUALITY: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate

LOW QUALITY: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate

VERY LOW QUALITY: Any estimate of effect is very uncertain

In keeping with other evidence rating systems used across the What Works Network, the 'high quality' rating in the GRADE approach is generally used for randomised trial evidence while evidence from sound observational studies would generally receive an initial rating of 'low quality'. However, the GRADE rating system allows flexibility in rating evidence at a higher or lower level depending on a range of considerations. For example, evidence initially rated as 'high' can be downgraded due to:

- Study limitations
- Inconsistency of results
- Indirectness of evidence
- Imprecision
- Reporting bias.

Similarly, evidence initially given a 'low quality' rating (such as evidence from observational studies) can be graded upwards if there is:

- A very large magnitude of effect
- A dose-response gradient; and
- All plausible biases would reduce an apparent treatment effect.

Alternatively, non-randomised studies can be upgraded to 'high' initial certainty if the ROBINS-I tool, primarily designed for cohort studies, has been used to assess the risk of bias in the studies.

As discussed in 11.2, risk of bias will be reported in the summary table to inform this judgement.

The GRADE Working Group website provides further information about the approach, links to publications where it has been applied, and tools for rating evidence review findings using the GRADE approach. The latest paper from the GRADE working group provides an update on applying the GRADE approach within 'complex systems'. Including:

- Considering complexity at the onset of the systematic review
- Choosing a threshold or a range that matches the needs of intended users of their review
- Assessing evidence from NRSs designs (cohort studies)
- Addressing sources of complexity when looking at the criteria within each GRADE domain for rating certainty

Thresholds for levels of certainty:
"'Certainty of evidence' is the confidence that the true effect of an intervention lies on one side of a specified threshold or within a chosen range" (Hultcrantz et al., 2017)
The newly proposed update from the GRADE working team illustrates different approaches for setting up the most appropriate thresholds or ranges for the certainty of evidence ratings depending on the contextualisation of the systematic review. Contextualisation refers to the purpose of the review and whether it is intended to inform a specific guideline/decision making process or not. They identified three main categories:

- Non-contextualised ratings are relevant for assessments conducted outside of a guideline
- Partly contextualised ratings are in the middle
- Fully contextualised ratings are relevant when systematic reviews are conducted as part of a specific guideline development or decision-making process

Further information can be found in (Annex 7).

A summary table of the proposed considerations and changes to the GRADE methodology can be found in (Annex 6).

13.2 Use of CERQual to rate the quality of evidence for findings in qualitative evidence reviews

The GRADE Working Group have also recognised the importance of assessing confidence in evidence from qualitative reviews and have developed CERQual (Confidence in the Evidence from Reviews of Qualitative research) to provide a transparent method of doing this. CERQual uses a similar approach conceptually to other GRADE tools, but is intended for findings from systematic reviews of qualitative evidence.

It is based on four components:

- Methodological limitations of the qualitative studies contributing to a review finding,
- Relevance to the review question of the studies contributing to a review finding,
- Coherence of the review finding, and
- Adequacy of data supporting a review finding.

When undertaking a qualitative evidence synthesis, the methodological limitations of each primary study included in the synthesis will be reviewed using the checklist in Appendix 2. Additionally, to assess the methodological limitations of the evidence underlying a review finding, review authors must make an overall judgement based on all of the primary studies contributing to the finding. This judgement needs to take into account each study's relative contribution to the evidence, the types of methodological limitations identified, and how those methodological limitations may impact on the specific finding.

Further information is available from the CERQual website and in a recent publication by Lewin et al (2015), Using Qualitative Evidence in Decision Making for Health and Social Interventions: An Approach to Assess Confidence in Findings from Qualitative Evidence Syntheses.

14. Making recommendations based on the evidence reviews

Each evidence review team will suggest recommendations for practice based on their findings, using the GRADE approach. This provides a clear and consistent approach to making recommendations based on the findings of evidence reviews. Additionally, each team will keep an evidence gap register and make recommendations about how gaps can be filled and where further research is required.

The evidence reviews and draft recommendations will be considered by the Centre's Advisory Panel and/ or round tables of experts who will provide comments and suggest possible refinements prior to publication.

Further information and links on developing recommendations in keeping with the GRADE approach can be found on the GRADE Working Group website.

15. Reporting Structure

We have not developed a template for the final report as each report is likely to be very different and flexibility here is more important than uniformity. Within this Methods Guide are the features that should be reported in each systematic review, but the relative importance of each will vary considerably from one systematic review to another.

16. Contact for questions or comments

If you have questions or would like to share your thoughts on our methods guide, please get in touch with the Evidence team at the What Works Centre for Wellbeing.

Email: evidence@whatworkswellbeing.org

17. Acknowledgements

We would like to thank the What Works Centre for Wellbeing Methods Group for the invaluable work they have done in compiling and assessing the advice in the Guide.

18. References

Bagnall AM, South J, Trigwell J, Kinsella K, White J, Harden A (2015) Community engagement – approaches to improve health: Map of the literature on current and emerging community engagement policy and practice in the UK. Leeds: Centre for Health Promotion Research, Institute for Health and Wellbeing, Leeds Beckett University.

Cochrane Collaboration (2009), Defining a Researchable Question: the PICOS Approach Cochrane Reviewers' Training Workshop January 22-23, 2009, slide share. Session Presenter: Marcus Vaska. Slides adapted from "Defining a Researchable Question." by Miranda Cumpston, with additions and deletions by Dr. Roger Thomas; "Review Protocol and Designing Your Research Question," by the Cochrane Collaboration

Critical Appraisal Skills Programme (CASP) 2014.CASP Checklists (qualitative checklist) Oxford. CASP

Dixon-Woods M, Shaw RL, Garwal SA, Smith JA,The Problem of Appraising Qualitative Research, *Qual Saf Health Care* 2004;**13**:223-225 doi:10.1136/qshc.2003.008714

Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011.

Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol 2017;87:4–13.

Lefebvre C, Manheimer E, Glanville J (2011) Searching for Studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0 (updated March 2011). The Cochrane Collaboration.

Lewin S, Glenton C, Munthe-Kaas H, Carlsen B, Colvin CJ, Gülmezoglu M, et al. (2015) Using Qualitative Evidence in Decision Making for Health and Social Interventions: An Approach to Assess Confidence in Findings from Qualitative Evidence Syntheses (GRADE-CERQual). PLoS Med 12(10): e1001895. doi:10.1371/journal.pmed.1001895

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *Preferred Reporting Items for Systematic Reviews and Meta-Analyses:* The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Early Intervention Foundation (2015) Translating the Evidence. A Brief Guide to the Early Intervention Foundation's Procedures for Identifying, Assessing, and Disseminating Information about Early Intervention Programmes and their Evidence.

National Institute for Health and Care Excellence (NICE) Methods for the Development of NICE Public Health Guidance (3rd edition), September 2012

Office for National Statistics, Self A, Measuring National Well-being: Insights across society, the economy, and the environment, 2014.

Ogilvie D, Fayter D, Petticrew M, Sowden A, Thomas S, Whitehead M, Worthy G, The harvest plot: A method for synthesising evidence about the differential effects of interventions
*BMC Medical Research Methodology*2008**8**:8, doi**:** 10.1186/1471-2288-8-8

Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, Hing C, Kwok CS, Pang C, Harvey I, Dissemination and publication of research findings: an updated review of related biases, Health Technol Assess. 2010 Feb;14(8):iii, ix-xi, 1-193. doi: 10.3310/hta14080.

Sutton A J, Duval S J, Tweedie R L, Abrams K R, Jones D R (2000) Empirical assessment of effect of publication bias on meta-analyses, BMJ 2000; 320:1574 doi: http://dx.doi.org/10.1136/bmj.320.7249.1574

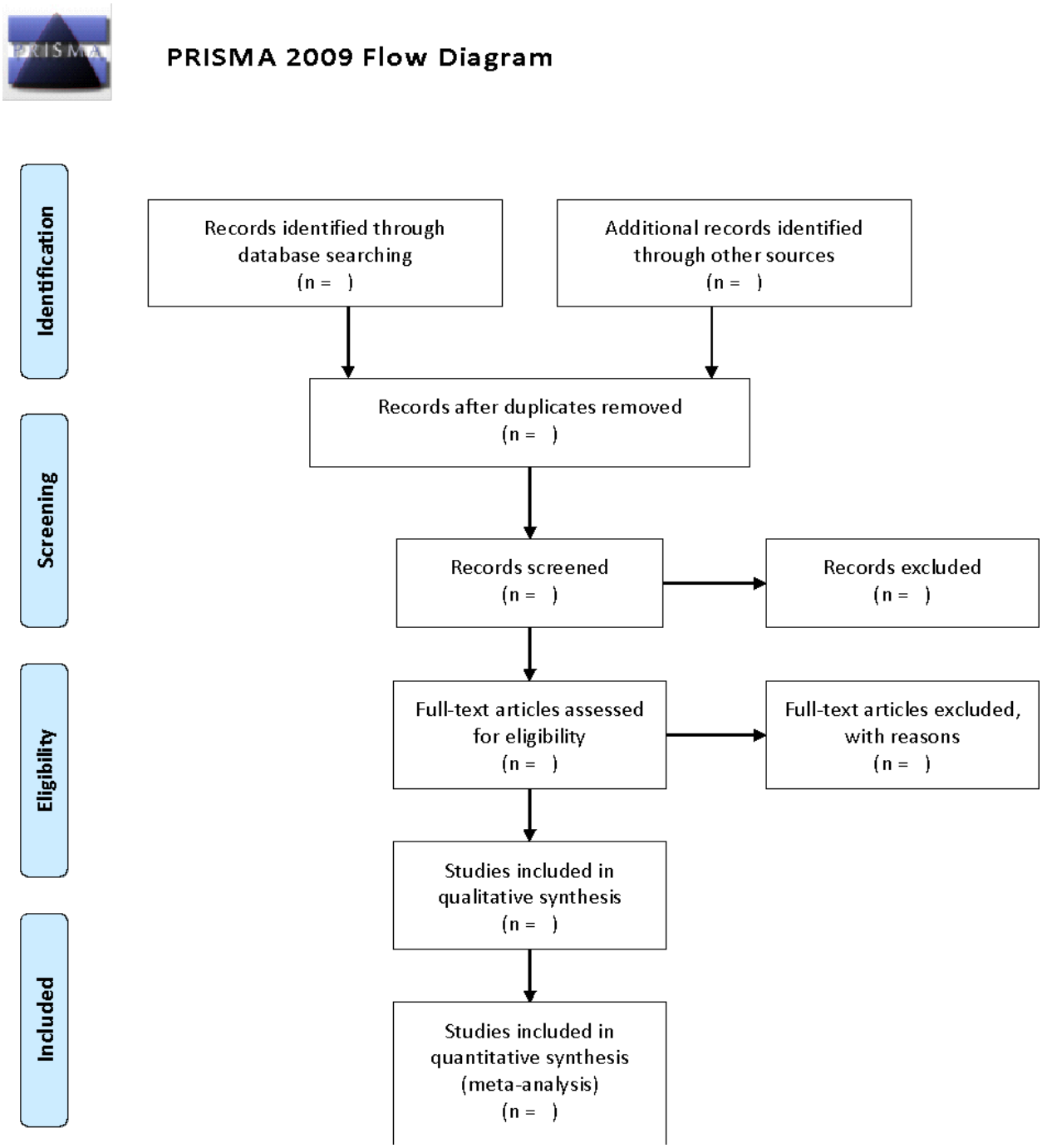
Systematic Reviews. CRD's guidance for undertaking reviews in health care. Centre for Reviews and Dissemination, University of York, 2009.

Welch, V; Petticrew, M; Petkovic, J; Moher, D; Waters, E; White, H; Tugwell, P; PRISMA-Equity Bellagio group (2015) Extending the PRISMA statement to equity-focused systematic reviews (PRISMA 2012): explanation and elaboration. Int J Equity Health, 14 (1). p. 92. ISSN 1475-9276

Montgomery P, Movsisyan A, Grant SP, et al. Considerations of complexity in rating certainty of evidence in systematic reviews: a primer on using the GRADE approach in global health. BMJ Glob Health 2019;4:e000848. doi:10.1136/bmjgh-2018-000848

Petticrew M, Knai C, Thomas J, et al. Implications of a complexity perspective for systematic reviews and guideline development in health decision making. BMJ Glob Health 2018;0:e000899.

Annex 1: PRISMA Flow Diagram



Annex 2: Quality checklist quantitative evidence of intervention effectiveness

Source: Based on Early Intervention Foundation Quality Checklist and amended for use.

How to use this checklist: This checklist is to be used to indicate if a specific study has been well designed, appropriately carried out and analysed, i.e. the confidence we can have in the results of whether an intervention was effective. This should be used for the summary table, to make brief comments on the risk of bias of each study (see section 11.2)
In turn, the overview of the study limitations will help to inform the quality of the overall body of evidence, as in section 13.

Evidence quality of intervention effectiveness / study limitations				
1. Was the evaluation well-designed?	**Yes**	**No**	**Can't tell**	**N/A**

• **Fidelity:** The extent to which the intervention was delivered with fidelity is clear - i.e. if there is a specific intervention which is being evaluated, this has been well reproduced.
• **Measurement**: The measures are appropriate for the intervention's anticipated outcomes and population.
• Participants completed the same set of measures once shortly before participating in the intervention and once again immediately afterwards
• An 'intent-to-treat' design was used, meaning that all participants recruited to the intervention participated in the pre/post measurement, regardless of whether or how much of the intervention they received, even if they dropped out of the intervention (this does not include dropping out of the study- which may then be regarded as missing data)
• **Counterfactual**:
• Assignment to the treatment and comparison group was at the appropriate level (e.g., individual, family, school, community)
• The comparison condition provides an appropriate counterfactual to the treatment group. Consider**:**
  ○ Participants were randomly assigned to the treatment and control group through the use of methods appropriate for the circumstances and target population OR sufficiently rigorous quasi-experimental methods (regression discontinuity, propensity score matching) were used to generate an appropriately comparable sample through non-random methods
  ○ The treatment and comparison conditions are thoroughly described.				

2. Was the study carried out appropriately? including appropriate sample	Yes	No	Can't tell	N/A
• **Representative:** The sample is representative of the intervention's target population in terms of age, demographics and level of need. The sample characteristics are clearly stated. • There is baseline equivalence between the treatment and comparison group participants on key demographic variables of interest to the study and baseline measures of outcomes (when feasible) • **Sample size**: The sample is sufficiently large to test for the desired impact. This depends most importantly on the effect size, however a suggestion could be e.g. a minimum of 20 participants have completed the measures at both time points within each study group. • **Attrition:** A minimum of 35% of the participants completed pre/ post measures. Overall study attrition is not higher than 65%. • The study had clear processes for determining and reporting drop-out and dose. Differences between study drop-outs and completers were reported if attrition was greater than 10%. • The study assessed and reported on overall and differential attrition • **Equivalence:** Risks for contamination of the comparison group and other confounding factors have been taken into account and controlled for in the analysis if possible: ○ Participants were blind to their assignment to the treatment and comparison group • There was consistent and equivalent measurement of the treatment and control groups at all points when measurement took place. • **Measures**: The measures used were valid and reliable. This means that the measure was standardised and validated independently of the study and the methods for standardization were published. Administrative data and observational measures may also have been used to measure programme impact, but sufficient information was given to determine their validity for doing this. • Measurement was independent of any measures used as part of the treatment. • In addition to any self-reported data (collected through the use of validated instruments), the study also included assessment information independent of the study participants (eg, an independent observer, administrative data, etc).				

3. Was analysis appropriate?	**Yes**	**No**	**Can't tell**	**N/A**
• The methods used to analyse results are appropriate given the data being analysed (categorical, ordinal, ratio/ parametric or non-parametric, etc) and the purpose of the analysis. • Appropriate methods have been used and reported for the treatment of missing data.				
4. Is the evidence consistent?				
• Are the findings made explicit? • Is there adequate discussion of the evidence both for and against the researcher's arguments? • Has the researcher discussed the credibility of their findings (e.g. triangulation, respondent validation, more than one analyst)? • Are the findings discussed in relation to the original research question?				

Annex 3: Quality checklist for qualitative studies (or qualitative components within mixed methods studies)

Drawing on the CASP approach, the following are the minimum criteria for inclusion of qualitative evidence in the review. If the answer to all of these questions is “yes”, the study can be included in the study in the review.

	Study inclusion checklist (screening questions)				
1. Is a qualitative methodology appropriate?		**Yes**	**No**	**Can’t tell**	**N/A**
Consider: Does the research seek to interpret or illuminate the actions and/or subjective experiences of research participants? Is qualitative research the right methodology for addressing the research goal?					
2. Is the research design appropriate for addressing the aims of the research? ***Consider:*** Has the researcher justified the research design (e.g. have they discussed how they decided which method to use)?					
3. Is there a clear statement of findings? ***Consider*:** Are the findings made explicit? Is there adequate discussion of the evidence both for and against the researcher’s arguments? Has the researcher discussed the credibility of their findings (e.g. triangulation, respondent validation, more than one analyst)? Are the findings discussed in relation to the original research question?					

The following criteria should be considered for each study to be included in the review (ie, those for which the answers to all of the screening questions were “yes”).

	Yes	No	Can’t tell
4. Was the data collected in a way that addressed the research issue? ***Consider:*** Is the setting for data collection justified? Is it is clear what methods were used to collect data? (e.g. focus group, semi-structured interview etc.)? Has the researcher justified the methods chosen? Has the researcher made the process of data collection explicit (e.g. for interview method, is there an indication of how interviews were conducted, or did they use a topic guide)? If methods were modified during the study, has the researcher explained how and why? Is the form of data clear (e.g. tape recordings, video material, notes etc)?			
5. Was the recruitment strategy appropriate to the aims of the research? ***Consider:*** Has the researcher explained how the participants were selected? Have they explained why the participants they selected were the most appropriate to provide access to the type of knowledge sought by the study? Is there are any discussion around recruitment and potential bias (e.g. why some people chose not to take part)? Is the selection of cases/ sampling strategy theoretically justified?			
6. Was the data analysis sufficiently rigorous? ***Consider:*** If there is an in-depth description of the analysis process? If thematic analysis is used, is it clear how the categories/themes were derived from the data? Does the researcher explain how the data presented were selected from the original sample to demonstrate the analysis process? Are sufficient data presented to support the findings? Were the findings grounded in/ supported by the data?			

Was there good breadth and/or depth achieved in the findings? To what extent are contradictory data taken into account? Are the data appropriately referenced (i.e. attributions to (anonymised) respondents)?			
7. Has the relationship between researcher and participants been adequately considered? ***Consider:*** Has the researcher critically examined their own role, potential bias and influence during (a) formulation of the research questions (b) data collection, including sample recruitment and choice of location? How has the researcher responded to events during the study and have they considered the implications of any changes in the research design?			
8. Have ethical issues been taken into consideration? ***Consider:*** Are there sufficient details of how the research was explained to participants for the reader to assess whether ethical standards were maintained? Has the researcher discussed issues raised by the study (e.g. issues around informed consent or confidentiality or how they have handled the effects of the study on the participants during and after the study)? Have they adequately discussed issues like informed consent and procedures in place to protect anonymity? Have the consequences of the research been considered i.e. raising expectations, changing behaviour? Has approval been sought from an ethics committee?			
9. Contribution of the research to wellbeing impact questions? ***Consider:*** Does the study make a contribution to existing knowledge or understanding of what works for wellbeing? e.g. are the findings considered in relation to current practice or policy?			

Annex 4: Quality checklist for economic evaluations (The Drummond Checklist, 1996)

Item		Yes	No	Not clear	Not appropriate
Study design					
1.	The research question is stated.				
2.	The economic importance of the research question is stated.				
3.	The viewpoint(s) of the analysis are clearly stated and justified.				
4.	The rationale for choosing alternative programmes or interventions compared is stated.				
5.	The alternatives being compared are clearly described.				
6.	The form of economic evaluation used is stated.				
7.	The choice of form of economic evaluation is justified in relation to the questions addressed.				
Data collection					
8.	The source(s) of effectiveness estimates used are stated.				
9.	Details of the design and results of effectiveness study are given (if based on a single study).				
10.	Details of the methods of synthesis or meta-analysis of estimates are given (if based on a synthesis of a number of effectiveness studies).				
11.	The primary outcome measure(s) for the economic evaluation are clearly stated.				
12.	Methods to value benefits are stated.				
13.	Details of the subjects from whom valuations were obtained were given.				
14.	Productivity changes (if included) are reported separately.				
15.	The relevance of productivity changes to the study question is discussed.				
16.	Quantities of resource use are reported separately from their unit costs.				

17.	Methods for the estimation of quantities and unit costs are described.				
18.	Currency and price data are recorded.				
19.	Details of currency of price adjustments for inflation or currency conversion are given.				
20.	Details of any model used are given.				
21.	The choice of model used and the key parameters on which it is based are justified.				
Analysis and interpretation of results					
22.	Time horizon of costs and benefits is stated.				
23.	The discount rate(s) is stated.				
24.	The choice of discount rate(s) is justified.				
25.	An explanation is given if costs and benefits are not discounted.				
26.	Details of statistical tests and confidence intervals are given for stochastic data.				
27.	The approach to sensitivity analysis is given.				
28.	The choice of variables for sensitivity analysis is justified.				
29.	The ranges over which the variables are varied are justified.				
30.	Relevant alternatives are compared.				
31.	Incremental analysis is reported.				
32.	Major outcomes are presented in a disaggregated as well as aggregated form.				
33.	The answer to the study question is given.				
34.	Conclusions follow from the data reported.				
35.	Conclusions are accompanied by the appropriate caveats.				

Source: Higgins, J and Green S (2011), Cochrane Handbook for Systematic Reviews of Interventions, The Cochrane Collaboration, version 5.1, section 15.

Annex 5: DRAFT 'Plain English' Guidelines

Evidence initially rated as 'high' can be downgraded due to:
- Study limitations
- Inconsistency of results
- Indirectness of evidence
- Imprecision
- Reporting bias.

Similarly, evidence initially given a 'low quality' rating (such as evidence from observational studies) can be graded upwards if there is:
- A very large magnitude of effect
- A dose-response gradient; and
- All plausible biases would reduce an apparent treatment effect.

Alternatively, it is possible to attribute a 'high' initial certainty rating for any type of study design providing that a rigorous tool has been used to assess risk of bias. The recognised tool to assess risk of bias for non randomised studies (NRSs) is (ROBINS-I) which is primarily designed for cohort studies.

GRADE / CerQual	**Icons**	**Studies (this will not be included - for our reference and discussion)**	**Description in plain English**
High quality High confidence		More than one high quality study with similar results, or one high quality 'upgraded'	**Strong evidence.** We can be confident that the evidence can be used to inform decisions.
Moderate quality Moderate confidence		Single high quality study with some limitations or multiple studies with some limitations	**Promising evidence.** Decision makers may wish to incorporate further information to inform decisions.
Low quality Low confidence		Single study with some limitations	**Initial evidence**. Decision makers may wish to incorporate further information to inform decisions.
Very low quality Very low confidence		It may be that studies show effects in different directions. Or, studies which have significant quality issues or may not be relevant.	**Unclear evidence.** Decision makers may not wish to act on this finding.

These statements should be supplemented by:
Strong, promising and initial evidence refer to high, moderate and low quality evidence / confidence as per GRADE and CERQual guidance. For further information on these classifications, please see [hyperlink to methods guide]. For further information on the underlying studies, please see [hyperlink to full reports]

All evidence should be considered alongside questions of possible benefits and risks, affordability, acceptability, feasibility and wider impacts, including equity issues, in the user setting. Where the evidence is less strong, these other considerations become even more important.
Further considerations have been added to Annex 6.

Annex 6: Rating certainty in systematic reviews of complex interventions using GRADE

Deciding on the scope of the review	
Use logic models to develop PICO and review questions	Logic models help in scoping, defining and conducting the review and in making the review relevant to policy and practice. Approaches have been developed to assist with this (Petticrew et al. 2019), (Rohwer et al., 2017), (Rehfuess et al., 2018), (Kneale et al., 2015).
Identify which tools to use to best describe the sources of complexity that users will require	There are several newly developed tools on using a complexity perspective in systematic reviews, such as the approach by Petticrew et al 2019, iCAT_SR, the CICI framework, TIDieR and PRISMACI
Using these tools identify contextual and implementation factors and other moderators of effect that may help explain heterogeneity and which will need separate GRADE certainty ratings	• In addition to the standard PICO question, identify in both the intervention and the system in which it is being used all the complexities and interactions that review users will want to know about. • Under intervention complexities, consider aspects of its implementation, such as theory of why and how the intervention is expected to work, the components, implementers, mediators, moderators, and causal pathways • Under system complexities, consider context, setting and any other independent interventions taking place
Defining thresholds or ranges for certainty of evidence ratings	
Define ‘certainty’ in a manner that matches the needs of the intended users of the review	• Decide among the three approaches to defining certainty of evidence: ‘non-contextualised’, ‘partly

	contextualised' and 'fully contextualised' ● In each case, specify the threshold or ranges used to rate certainty of evidence ● For 'non-contextualised' reviews, consider the utility of using GRADE for the 'non-null' effect
Rating certainty of evidence using GRADE	
Initially rate any body of evidence as 'high' if a rigorous tool is used to assess risk of bias in NRSs (ie, ROBINS-I), otherwise, use the 'standard' GRADE guidance	Consider using Cochrane Risk of Bias (RoB V.2.0) tool for randomised controlled trials Consider using ROBINS-I for cohort-type studies
Give extra scrutiny to the impact of lack of blinding providers/participants on overall risk of bias for outcomes	If lack of blinding of either participants of providers is unlikely to affect assessment of outcome (such as when using objective outcome measures, for example, mortality), then consider not downgrading evidence for lack of blinding for that outcome.
Consider the effect of bias associated with deviation from the intended intervention	● Deviations, such as poor adherence, poor implementation and cointerventions in relation to the effect of starting and adhering to an intervention, may lead to bias and may be downgraded by one level ● Consider not downgrading if assessing the effect of assignment to the intervention, when deviations do not occur in relation to usual practice and groups remain balanced
Consider multiple criteria for judging inconsistency of evidence	Assessment of heterogeneity should always start off with an appraisal of study heterogeneity, including heterogeneity in PICO elements as well as methodological aspects ● Assessment of heterogeneity should take account of multiple rather than single criteria for inconsistency (eg, I2 and its p value, overlap of CIs and degree of variation within chosen thresholds) ● Consider whether definition of certainty of evidence influences nature of inconsistency assessment (eg, when effect sizes across all studies

	are consistently in the same direction outside of the null effect or a given threshold of interest, then downgrading for inconsistency is not warranted despite other measures) ● Consider different analytical methods to explain heterogeneity (eg, subgroup analysis, metaregression and qualitative comparative analysis)
Rate imprecision of evidence with regard to the adopted definition of 'certainty'	Consider whether definition of certainty of evidence influences nature of imprecision assessment ● For 'non-contextualised' systematic reviews definition, a certainty that the effect lies within estimated CIs or prediction intervals, a GRADE assessment for imprecision can usually be omitted as assessment of precision is dependent on the chosen range ● For 'partly contextualised' systematic reviews, consider whether the point estimate would represent a trivial, small, moderate or large absolute effect ● For 'fully contextualised' systematic reviews, simultaneously consider all important outcomes to determine precision of the effect estimate
Examine indirectness of evidence by way of assessing important differences in the evidence base beyond what is expected	● Consider grouping studies, synthesising evidence and rating certainty in the estimates of effect for separate outcomes according to the relevant sources of complexity identified at the start of the review ● Consider splitting the questions to answer subset conditions, downgrading only for those with less certain evidence. Do not downgrade for indirectness if observed differences are unlikely to affect the outcome
Consider publication bias	● Conduct extensive grey literature searches and expert contacts to identify reports and working papers ● Consider sponsorship of studies by any vested industries as well as

	potential 'allegiance bias'
Upgrading evidence	• Consider upgrading certainty of evidence for a dose–response relationship related to the level of implementation • Consider upgrading evidence for a body of evidence from studies with low implementation fidelity positive results which counteract plausible residual bias or confounding
Use logic models to investigate coherence of evidence across the causal pathway	• Consider assessing the coherence of evidence across different links in the causal pathway at the end of evidence synthesis. This judgement should be made outside of the GRADE framework

Source: Montgomery P, Movsisyan A, Grant SP, et al. (2019) Considerations of complexity in rating certainty of evidence in systematic reviews: a primer on using the GRADE approach in global health. BMJ Glob Health.

Annex 7: Approaches for setting thresholds using GRADE

Setting	Contextualisation	Threshold or range	How to set	What certainty rating represents
Primarily for systematic reviews and health technology assessment	Non-contextualised	Range: 95% CI OR≠1; RR≠1; HR≠1; RD≠0	Using existing limits of the 95% CI, which implies that precision is not routinely part of the rating Using the threshold of null effect	Certainty that the effect lies within the CI Certainty that the effect of one treatment differs from another
Primarily for systematic reviews and health technology assessment	Partly contextualised	Specified magnitude of effect	For example, small effect is the effect small enough to not use the intervention if adverse effects/costs are appreciable	Certainty in a specified magnitude of effect for one outcome (eg, trivial, small, moderate or large)
Primarily for	Fully	Threshold	Considering the	Confidence that

practice guidelines	contextualised	determined with consideration of all critical outcomes	range of effects on all critical outcomes, and the values & preferences for those ranges	the direction of the net effect will not differ from one end of the certainty range to the other

RD, risk difference; RR, risk ratio.

Source: Montgomery P, Movsisyan A, Grant SP, et al. (2019) Considerations of complexity in rating certainty of evidence in systematic reviews: a primer on using the GRADE approach in global health. BMJ Glob Health.

Funded by:

E·S·R·C