EVIDENCE STANDARDS: A DIMENSIONS OF DIFFERENCE FRAMEWORK FOR APPRAISING JUSTIFIABLE EVIDENCE CLAIMS[1]

David Gough, EPPI-Centre, SSRU, UCL Institute of Education, University College London

INTRODUCTION

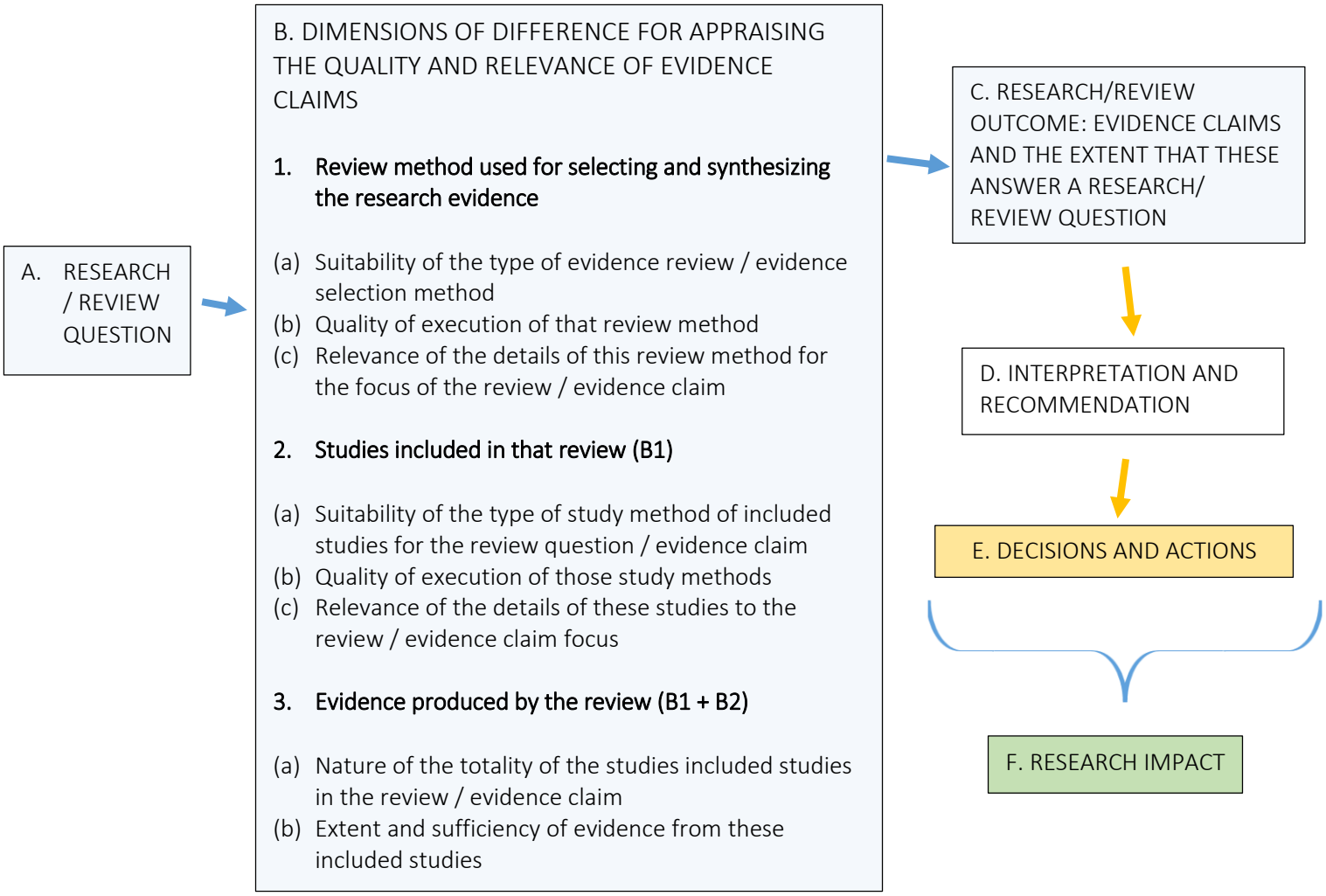
When people undertake research and produce findings, they often make an evidence claim about what the findings do or do not show. The evidence claim may be made on the basis of a body of work or on the basis of an individual study. An example of a body or work would be the combined work of a research team, another example would be a systematic review of research addressing a particular research question. A systematic review is 'a review of existing research using explicit, accountable rigorous research methods' (Gough et al, forthcoming).

This short paper provides an overarching conceptual framework for considering the main dimensions involved in appraising and thus justifying such evidence claims (whatever the research question or research method or evidence claim). These 'dimensions of difference of justifiable evidence claims' are presented as one part (Component B) of a wider framework of the research ecosystem (ranging from research questions to evidence claims and the outcomes and uses of such claims). The broader framework includes the following six main components (see Figure 1).

A. The research question – what question is being asked of the research evidence.
B. The 'dimensions of difference of appraising evidence claims'. This distinguishes three dimensions involved in appraising the quality and relevance of evidence claims. Each of these dimensions has sub-dimensions as listed in Figure 1 and later in the text of this paper.
    1. Review method used for selecting and synthesizing studies (which may or may not be systematic)
    2. Studies included in that review (B1)
    3. Evidence produced by the review (B1 + B2).
C. The review outcome – the evidence claims being made from the research and the extent that these address the research question being asked. Including the limits of the claim which may involve: (i) Certainty; (ii) Extent of applicability.
D. Interpretation and recommendations - arising from the evidence claim but that involve further information
E. Decisions and actions – that may arise from considering the interpreted evidence claims
F. Contribution of the evidence claim to an academic field or discipline or to wider society

[1] Please cite as: Gough DA (2016) *Evidence Standards: A Dimensions of Difference Framework for Appraising Justifiable Evidence Claims* London: EPPI-Centre, UCL.

Figure 1: A dimensions of difference framework for appraising evidence claims



These components are discussed briefly in turn with most attention being given to Component 'B', the dimensions of difference of making justifiable evidence claims.

Many tools exist to appraise different aspects of the quality and relevance of studies (and reviews of those studies) and the evidence claims that are made about them. This framework is not such a quality and relevance appraisal tool. It does not provide criteria for considering how different types of research studies might make or test the strength of different types of evidence claims. It is instead a framework to understand the different types of appraisal judgements that are being made. The framework thus provides a way to understand the role that such more specific tools play within a wider conceptualization of the process of quality and relevance appraisal in making (or checking) evidence claims. The text in this paper refers predominantly to evidence from research studies but the logic can also be applied to the selection and appraisal of other types of evidence.

THE FRAMEWORK

A. RESEARCH / REVIEW QUESTION

Research is driven by research questions. The research aims to answer those questions leading to evidence claims based on the research findings. There are very many different types of research questions and a diverse range of research methods developed to address them. The logic of clarifying the basis for making an evidence claim applies to all such research questions and research methods – ranging from experimentally controlled trials that aim to assess the impact of interventions to small exploratory studies that aim to develop theories and concepts.

Research questions can vary substantially on the same topic due to different perspectives (values and priorities) of those included in asking the questions. Research questions can be appraised on the extent that they ask relevant, ethical and equitable questions and how different perspectives were involved in their creation.

B. DIMENSIONS OF DIFFERENCE OF APPRAISING EVIDENCE CLAIMS

In addition to the appropriateness of research questions, there is the issue of the fitness for purpose of the methods that are applied to address these questions and of the quality and relevance of the resultant evidence claims.

The main ways in which evidence claims are appraised (in terms of them being justifiable by the research evidence) can be characterized as falling into three related dimensions. (This does not of course include all the other ways in which people might agree or disagree with research in terms of whether it asks relevant questions Component A). These three dimensions are:

B1. Review method: how the evidence on which the evidence claim is being made were selected and synthesized
B2. Included studies: the nature of the studies from which the evidence came
B3. Evidence produced: the nature, extent and sufficiency of this evidence

These three dimensions and their sub-dimensions are described in turn.

Dimension B1. Review (evidence selection) method used for selecting and synthesizing the research findings on which an evidence claim is made

If an evidence clam is being made then it is important to know how the evidence was selected and analysed. Normally you would expect that that strong evidence claims would arise from some sort of consideration (review) of what research evidence can be drawn upon rather than based upon one research study that happened to be at hand. In some cases, a consideration of the available relevant evidence may lead to evidence claims on the basis of a single study; for example a high quality total population survey on the prevalence of a phenomena. In other cases, a number of different studies may be found and analysed. For clarity, this framework will assume that an evidence claim is being made on the basis of

some form of review of available research evidence. Hence the term 'review' is used to refer to any collection of studies used to make an evidence claim.

In the last twenty years there has been increasing concern to be clearer about what is already known about a research question from the findings of pre-existing studies. There has also been increased concern that the methods for identifying and analysing the results of such pre-existing studies should be systematic and rigorous and explicit just as is expected for primary research. The term 'systematic review' has been developed to describe reviews with such formal methods (Gough et al 2012, forthcoming).

Whatever research is being considered, we need to appraise how it was undertaken in order to decide whether its findings are trustworthy and relevant. Being systematic in method does not necessarily mean that the findings will be fit for purpose and that justifiable evidence claims can be made. It may be that the method is not appropriate, that the method is badly executed, or that the focus of the study was not relevant to the research question. This provides the following sub-dimensions for appraising how research evidence is selected and analysed to make an evidence claim:

Sub-Dimension B1a: Suitability of the type of review / evidence selection method used
There are many different formal and informal methods for reviewing research evidence relevant to answering a research question (see Gough et al 2012, forthcoming). The suitability of review method will depend on the nature of the review question/ evidence claim. If, for example, the review question/evidence claim is about the relative impact of two contrasting interventions then a simple statistical meta-analysis systematic review may be a fit for purpose way to select and analyse the findings of included studies. If, on the other hand, the question is about process and complexity then a different type of study selection and review may be appropriate.

Sub-Dimension B1b: Quality of execution of that review / evidence selection method
Whatever method is used to review the evidence, the method could be applied well or poorly according to the accepted rules for applying that method. There are methodological principles that apply across all types of research but there are also varying quality issues specific to particular research paradigms, questions and methods. Methodological issues in a statistical meta-analysis to answer an impact question (using a priori paradigms) are, for example, very different to those in a thematic synthesis (using more iterative inductive methods) to address a conceptual question. It is also important to note that knowledge of the quality of execution of a study is dependent on the quality of reporting, (though poor reporting maybe an indirect indicator of poor standards more generally).

Sub-Dimension B1c: Relevance of the details of the review method for the focus of the review / evidence claim
However fit for purpose the method of review of evidence (B1a), and however well executed this method (B1b), the focus of the review may vary in its relevance to the focus of the review question and evidence claim. It may be, for example, that the definition (inclusion criteria) of studies to be included in the review are not fully fit for purpose for the review question and evidence claim; for example, the breadth of the review question might

be broader than the inclusion criteria of studies to answer that question. The consideration of relevance may also include value issues such as the appraisal of ethical and equity issues in how the review was conducted.

Tools for reporting and appraising methods for reviewing studies

There are many reporting guidelines and standards to advise on how to write up some of all of these components of methods of review. PRISMA[2] and Rameses[3], for example, provide guidance on reporting of methods for certain types of systematic review. Reporting standards are important in helping to achieve consistency across papers but they are also a statement of what is important methodologically and so can be used to appraise methodological adequacy.

Other guidelines more specifically focus on appraising review methods such as AMSTAR[4].

As these guidelines and tools tend to be concerned with particular review questions, they tend to focus on adequacy of the execution (Component B1b) of that method rather than on the suitability of the method (Component B1a). Some tools, such as ROBIS[5] also focus on the relevance of the details of the method for the review question (Component B1c).

Dimension B2. The studies / evidence sources selected and included

Dimension B2 is concerned with the individual studies being used in a review. Although the method of review (B1) will determine what studies are included, B2 is concerned with the resultant studies that are included.

The sub-dimensions for appraising the quality and relevance of individual studies / sources of evidence (Dimension B2) are essentially the same as for the appraising review methods (Dimension B1) but applied to the included studies rather than the review. These sub-dimensions are: (a) Suitability of the methods used by the individual included studies; (b) Quality of execution of those methods by these individual studies; (c) Relevance of the details of the studies' methods for the focus of the review and evidence claim (Gough 2007).

Sub-Dimension B2a: Suitability of the type of methods used in the included studies for the review question / evidence claim

The type of methods used by the research studies included in the review will impact upon the nature of evidence available in the review. The suitability of method will depend on the nature of the review question/ evidence claim. A simple randomized controlled trial may, for example, be a systematic and rigorous way to address simple questions of relative impact of an intervention, but may not be so powerful in assessing causal processes. Similarly, some review questions /evidence claims may require multiple review /evidence collection strategies.

[2] http://www.prisma-statement.org/

[3] http://www.ramesesproject.org/Home_Page.php#rameses1

[4] https://amstar.ca/

[5] http://www.bristol.ac.uk/population-health-sciences/projects/robis/

Sub-Dimension B2b: Quality of execution of those methods

Whatever methods were used by the included research studies, the methods could be applied well or poorly according to the accepted rules for applying that method. As with the quality of execution of review (Sub-dimension B1b), the expectations of methods will vary. The expectations for a randomized controlled trial to address an impact question (with an a priori paradigm and concerns about avoiding bias) will, for example, be very different to the expectations for a qualitative case study (with an iterative paradigm and concerns that the studies reflected the phenomena they are trying to conceptualize). Also similarly, it may be difficult to distinguish issues of quality of execution from qualities of reporting.

Sub-Dimension B2c: Relevance of the details of the studies' methods for the focus of the review / evidence claim

However fit for purpose the type of methods used by the included studies (B2a), and however well executed these methods (B2b), the focus of the individual studies may vary in their relevance to the focus of the review question / evidence claim. It may be that, for example, the sample and population or the research measures are not that relevant to the evidence claim. Again, the consideration of relevance of focus may also include appraisal of ethical and equity issues in the research.

Tools for reporting and appraising methods of included studies

There are many reporting guidelines and appraisal tools for assessing some or all of these sub-dimensions of the quality and relevance of research studies. Most of the methods are research method specific and so are concerned more with execution (Sub-dimension B2b) rather than suitability of method (Sub-dimension B2a). Well known tools for appraising included studies in systematic reviews include GRADE[6] for a range of quantitative research methods and GRADE-CERQual[7] for qualitative research studies. Appraisal tools for review methods such as AMSTAR also consider the appraisal of included studies as part of the review process.

Dimension B3. Evidence produced by these review / evidence selection methods

In addition to the method of reviewing evidence (B1) and appraising included studies (B2), there is the issue of the quality and relevance of the total evidence produced by the review that is considered in order to make an evidence claim. This dimension of an evidence claim has two sub-dimensions: B3a, the nature of the totality of included studies; and B3b, the extent of the evidence that they provide.

Sub-Dimension B3a: Nature of the totality of included studies

Whatever the methods for reviewing studies and the methods of the included studies, there are quality and relevance issues arising from across the total sample of included studies. These can include, for example, issues of publication bias, heterogeneity and coherence across the study findings.

[6] http://www.gradeworkinggroup.org/

[7] http://www.cerqual.org/

Sub-Dimension B3b: Extent and sufficiency of evidence in the review

Reviewers of evidence are dependent on what research evidence is available. Even if fit for purpose methods of review and high quality and relevant included studies are used, this does not necessarily mean that the review will produce sufficient good quality and relevant evidence to make a strong justifiable evidence claims. The relevant evidence may not be available when sufficient relevant studies have been:

(i) Not undertaken;
(ii) Undertaken but not reported (publication bias);
(iii) Undertaken and reported but not identified by the review (a failure of the review on Dimension B1 in this framework);
(iv) Not properly managed by the review (a failure of the review on Dimension B1 in this framework);
(v) Not of appropriate quality and relevance to provide meaningful evidence (a failure of the included studies on Dimension B2 in this framework);
(vi) Not able to obtain strong evidence (even when these studies have been undertaken and reviewed appropriately).

Tools for appraising the totality of evidence

Some appraisal tools such as GRADE and GRADE-CERQual appraise included studies as part of a broader appraisal of the evidence provided by a review. They therefore consider both included studies (Dimension B2) and the totality of evidence (Dimension B3) (see Liabo et al forthcoming).

C. REVIEW OUTCOME: EVIDENCE CLAIMS AND THE EXTENT THAT THESE ANSWER THE REVIEW QUESTION

The findings of a study, a group of studies or a systematic review of studies provide a statement of the evidence in relation to a research question. The nature of the claim will depend on the nature of the question being asked and may be a probabilistic statement of the certainty of something. Evidence claims may make statements about the limits of the applicability of the claim, Evidence claims therefore vary in both the strength of the claim, its applicability, and how justifiable it is. A low level of claim and a narrowly applicable claim may both be more easily justified by the research evidence.

The quality and relevance of the method of all the three dimensions of Dimension B in the framework (B1 Evidence selection and analysis; B2 Included studies; and B3 Totality of evidence) all contribute to the assessment of the strength of evidence claim. They do this in terms of both: (i) the appraisal of quality and relevance of each sub-dimension in relation to the claim; and (ii) how the different dimensions and sub-dimensions combine and cohere together to provide a relatively weak or strong evidence claim.

The framework described in this paper assume that an evidence claim is based on a coherent relationship between review questions, review methods, included studies, and totality of evidence produced. It is, however, possible for there to be justifiable evidence claims from studies even if they do not relate directly to the review authors' (or authors of

included studies) original research questions. Others may examine the evidence and make justifiable evidence claims on the basis of the evidence found.

Further comments about tools for appraising evidence claims

Many well-known tools only appraise certain dimensions or sub-dimensions (see tables in Liabo et al forthcoming). Some tools such as PRISMA, AMSTAR and AMSTAR are concerned in different ways with review methods, many tools are only concerned with quality of execution of specific methods in primary research studies. Tools such as GRADE and GRADE-CERQual cover many aspect of both included studies (B2) and the totality of evidence (B3) but not the review method (B1).

Some tools provide scoring systems based on adding up the number of items that meet certain standards. This can have the disadvantages of assuming that each item has the same weight in a scoring system. They may therefore allow evidence with a few critical flaws but otherwise appropriate methods to obtain misleadingly high quality scores. Systems such as GRADE therefore use a different approach by starting with an initial good score and then increasing or reducing the score if there are reasons (from any of the questions that the tool asks) to indicate that there are bases for any such raising or lowering of the quality and relevance score.

D. INTERPRETATION AND RECOMMENDATION

In order to move from evidence claims to recommendations requires interpretation and decisions about application. This may involve taking into consideration many other possible types of information (beyond the research evidence in the evidence claim) including perspectives, politics, laws, resources, and broader contextual factors that may affect the impact of any decision informed by the evidence claim (Gough 2006). These other sources of information may themselves be appraised on the basis of their evidence claims whether or not they are also forms of research.

The GRADE system provides a formal system for considering particular types of other evidence in order to move from the evidence claims to recommendations about health interventions. These include: risks of using or not using an intervention; the burden of experiencing the intervention; the financial costs; and social values and preferences of those involved. The DECIDE[8] project took this further to provide a structured framework for considering a range of factors that could be considered in making recommendations from research (Seriousness of the problem; Number of people affected; Quality of evidence; Size of benefits and adverse effects; Resource use; Value for money; Impacts on equity; Feasibility; Acceptability).

There are also many other approaches for interpreting evidence that vary in formality and the involvement of different perspectives in a deliberative process. The National Institute for Health and Care Excellence[9] (NICE) in England, for example, has a stakeholder committee to deliberate about the interpretation of evidence claims from systematic reviews using

[8] http://www.decide-collaboration.eu/key-decide-tools
[9] https://www.nice.org.uk/

other evidence (including practitioner and user perspectives and experience) to fill evidence gaps and provide a context for interpreting evidence for policy and practice contexts to develop national guidance on health and social care services.

E. DECISIONS AND ACTIONS

Interpretations and recommendations from evidence claims are of course not the same as decisions (which may or may not be informed by research evidence as well as many other factors) (Gough 2006).

Consideration of one or more evidence claims in making a decision can also be distinguished from implementation strategies. Deciding on the basis of an evidence claim that, for example, doctors should wash their hands frequently, can be different to strategies to enable the implementation of that decision and change doctors' hand washing behaviour.

F. RESEARCH IMPACT

Finally, an evidence claim and its interpretation and decisions informed by such evidence claims can be distinguished from the impact that the evidence claim is perceived to make. Research may have many forms of instrumental and enlightenment impact on individuals or groups. Impact can also be appraised in terms of the contribution that individual or groups of evidence claims are thought to make to academia or to wider society. Various criteria can be developed to assess the nature and extent of different types of impact of different evidence claims but the strength of the evidence claim is unlikely to the only relevant factor. There is also the extent or breadth of an evidence claim and its relevance to different forms of impact as well as all the many other social and economic factors that influence research impact.

REFERENCES

Gough D (2006) User led research synthesis: a participative approach to driving research agendas. *Presented at: Sixth International Campbell Colloquium*, Los Angeles, 22–24 February. Available at: https://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/User%20led%20research%20synthesis%20new%20logo.pdf

Gough D (2007) Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, 22 (2): 213-228.

Gough D, Thomas J, Oliver S (2012) Clarifying differences between review designs and methods, *Systematic* Reviews, 1:28Gough D, Oliver S, Thomas J (forthcoming 2017) *Introduction to Systematic Reviews 2nd Edition.* London; Sage.

Liabo K, Gough D, Harden A (forthcoming 2017). Developing justifiable evidence claims. In Gough D, Oliver S, Thomas J (Eds.) *Introduction to Systematic Reviews 2nd Edition*. London; Sage.