Chapter 1

Appraising Evidence Claims

David Gough
EPPI-Centre, University College London

For research evidence to inform decision making, an appraisal needs to be made of whether the claims are justified and whether they are useful to the decisions being made. This chapter provides a high level framework of core issues relevant to appraising the "fitness for purpose" of evidence claims. The framework includes (I) the variation in the nature of research, the evidence claims it produces, and in the values, perspectives, and ethical issues that underlie it; (II) the main components of the bases of evidence claims in terms of (i) how relevant evidence has been identified and synthesized to make a claim, (ii) the technical quality and relevance of the included evidence, and (iii) the totality of evidence to justify the warrant of the evidence claim (including the potential for there to be alternative explanations); (III) evidence standards to appraise evidence claims and examples of guides and tools to assist with aspects of such appraisal; and (IV) engagement with evidence: (i) the communication of evidence claims, (ii) the fitness for purpose of these evidence claims for decision makers, and (iii) and the interpretation of such claims to provide recommendations and guidance.

Over the past 10 years there have been many calls for better use to be made of the findings of research in policy, practice, and personal decision making. This raises the issue of how research findings are to be assessed.

Whatever the research questions and method and whoever instigated the research to be undertaken, studies aim to reach conclusions based on the research and so make some form of "evidence claim." How can potential users of research be sure that any such evidence claim is "fit for purpose" in being both trustworthy and relevant to their needs? This chapter does not discuss detailed methodological issues. Instead it provides a high-level framework for considering these issues of evidence use within an evidence ecosystem (summarized in Table 1).

Review of Research in Education
March 2021, Vol. 45, pp. 1–26
DOI: 10.3102/0091732X20985072

TABLE 1
Fitness for Purpose of an Evidence Claim Framework

Rigor, Explicitness and Accountability, Coherence, Consistency, Appropriateness, and Relevance of the Methods Used to Make and Interpret an Evidence Claim
I. The evidence claim
i. Perspectives underlying the claim (values, priorities and theoretical constructs), how these are determined, and their fit with research users' needs
ii. Nature, scope, level, focus, generalizability, certainty, and alternative explanations for the claim
II. Basis for the evidence claim
1. What were the review methods (for bringing together the relevant evidence base)?
i. Technical quality of the execution of the review method
ii. Appropriateness of the review method
iii. Relevance of focus of how the review method was applied
2. How were included studies appraised?
i. Technical quality of the execution of the methods of the included studies
ii. Appropriateness of these research methods
iii. Relevance of focus of how these methods were applied (including ethics of the process by which the research was undertaken)
3. What was the resulting totality of evidence used to make the evidence claim?
i. Nature of evidence
ii. Extent of evidence
iii. Does this evidence justify (warrant) the evidence claim?
III. Appraisal of evidence claims: evidence standards, tools, and guides
1. What are the evidence standards for appraising the warrant?
2. What methods, guides, and tools are used to apply these standards
i. Do these consider all of the relevant technical issues for the evidence claim being made?
ii. Are they themselves technically adequate and appropriate?
iii. Are there still dangers of being misled by the narrowness of any appraisal?
IV. Engaging with evidence
1. Communication: what is the nature of the claims and their warrants
2. Fitness for purpose: of the evidence claims with the users' needs
3. Recommendations and guidance from evidence
i. Is the guidance relevant to the decision?
ii. What other information and perspectives were used to develop the guidance?
iii. Was the process for making these recommendations rigorous, explicit, accountable, and appropriate?

Section I of the chapter considers the nature of an evidence claim: first, how evidence is produced and interpreted from different values and perspectives and including any ethical issues, and second, the nature of the claim being made and its scope, level, focus, generalizability, and certainty.

Section II examines the evidence base on which a claim is made. First, an evidence claim needs to be appraised against all of the available relevant evidence, so how is all of the relevant evidence identified and brought together to make the claim? Second, the technical quality and relevance of that evidence base. This includes the quality and relevance of: the methods by which the evidence is identified, the studies included, and the totality of evidence produced.

Section III discusses variations in evidence standards as well as guides and tools for producing, reporting, and appraising evidence claims. It provides some examples of such guides and tools and locates their position within Section II of the framework.

Section IV introduces some aspects of engagement with evidence. It provides examples of the ways that evidence claims are communicated (including web-based evidence portals), the assessment of the fitness for purpose of such claims for decision makers, and recommendations and guidance for practice that are informed by evidence claims.

The chapter concludes with summary questions about appraising the fitness for purpose of evidence claims arising from these four components of the framework. Critical appraisal can therefore be seen to include assessments of the following: the methods to bring together and review the evidence, the methods of the studies included in such a review, the nature and extent of the totality of evidence then used to support an evidence claim, and the fitness for purpose of the claim for specific uses. The chapter is written for those who use research evidence and those who communicate and broker the use of evidence to navigate the evidence claims that they may encounter. The chapter is also relevant to producers of research in considering the nature of the claims that they make.

The author is an academic researcher studying the use of research and variations in methods of systematic review. He sees the use of research evidence as occurring within an evidence ecosystem (Best & Holmes, 2010, Gough et al., 2019) and that the two core parts of that system are, first, the perspectives and values with which we use and create research evidence, and second, the formal accountable processes for gathering together what we know (and don't know) from research and for interpreting its relevance for decision making.

THE EVIDENCE CLAIM

Research is systematic enquiry and provides a special form of information to enable decision making (Stenhouse, 1981). For the special nature of this information to be appropriately used, clarity is required on the nature of the claim, the perspectives that framed it, and the trustworthiness and relevance of the claim, including alternative explanations for the data used as warrant for the claim.

Research evidence is very diverse and driven by many different perspectives, research questions, and methods, which makes it challenging to evaluate evidence claims. A first stage in assessing whether an evidence claim is justified and fit for the purposes to which it will be applied requires clarification of the variation in the

- values and perspectives framing the research,
- consequent research methods used, and
- resultant findings and evidence claims.

Perspectives Driving the Interpretation and Production of Research and the Evidence Claims Made

People vary in how they perceive the world and what is important to them. They vary in their values, the research questions that they ask, and the research findings that they might consider. The planning, production, and interpretation of research is thus driven by people's implicit and explicit perspectives (ideological and theoretical assumptions, values, and priorities).

There is variation in the extent that there is consensus in worldviews and pertinent "facts" within and across cultures and historical time. Individuals and groups vary in what aspects of an issue they consider important, how these should be best analyzed, and what success means. Different perspectives may consider different "facts" and use methodologically robust methods to reach different valid conclusions. Disagreements about research evidence may be due to differences in the perspectives underlying the production of that evidence, or due to issues of methodological robustness, or to a mixture of the two. Being explicit about both perspectives and methods can help reduce hidden bias and misunderstandings about research findings.

In practice, much of the research agenda (and thus what is produced by research) is driven by the perspectives and interests of those funding and undertaking research. In recent years there has been growing concern that there should be greater societal involvement in the prioritization of research questions and the interpretation of research findings. This is partly a broad issue of societal engagement in research and also a more specific democratic issue of a right to involvement by those likely to be affected by the research (Djulbegovic & Guyatt, 2017; Gough & Elbourne, 2002; Oliver et al., 2018).

There are also pragmatic reasons. First, involving broader perspectives in the research process may lead to overt consideration of different ways of framing an issue and lead to more relevant and useful research. Second, the engagement of these other actors in the research process may make it more likely that these individuals and groups are aware of and motivated to make use of research in their own thinking and decision making. All these different factors have led health research to involve patients with carers and physicians in setting research agendas and in the interpretation of research findings (e.g., the James Lind Alliance[1]). This has not been so evident in education research where there is relatively little input from parents or school students as users of the services (Gough et al., 2018).

The use of research findings is sometimes seen as a one-way process of research production leading to translation, evidence-informed decision making, and then implementation of those decisions. Use of research is now often understood as arising from a more dynamic interactive two-way process of "pull" (demand by users of research including those affected by the study findings) as well as "push" (the production and translation of research) within an evidence ecosystem (Best & Holmes, 2010; Gough et al., 2019). It is therefore important to consider the perspectives that drive both the use and the production of research.

In education research, much of the evidence (particularly that communicated through web-based portals to users of research) is concerned with questions of the effectiveness of interventions and the fidelity of their implementation. The interest of intervention developers and evaluators is concerned with the range of competing intervention programs to add on to school provision. Providers and users of services on the other hand would be expected to be problem driven with an interest in the range of options that could be considered for responding to educational challenges. Research that examined theories of change and mechanisms involved with different educational strategies and how research evidence may inform the addition of innovative approaches that are adapted to local contexts could be particularly useful. An appraisal of research evidence is thus not only about whether an evidence claim is justifiable in its own terms. It is also about whether it is addressing the most appropriate questions for different users of research. If most research has arisen from a particular perspective such as the efficacy of "add-on" programs, then this may limit the options for those wishing to apply broader question-driven approaches to using research to inform their policy and practice.

Perspectives also link to research ethics. These are the moral values that cover the behavior of researchers. The American Educational Research Association (AERA)'s Code of Ethics (2011) lists five principles and 22 ethical standards for education researchers. They range from principles of social responsibility to ethical standards of competence, use and misuse of expertise, and avoidance of harm. Concerns about principles and standards can include morality of the researcher behavior, participation of users in a democratic research process (Gough & Elbourne, 2002), issues of equity (Welch et al., 2015), and issues of research waste (Chalmers & Glasziou, 2009). If a research study is not considered ethical, then this may also undermine the evidence claims that it makes.

In discussions about research evidence, the emphasis is often on technical issues of methods rather than on the perspectives that frame the research questions and the interpretation of research or examining alternative framings and explanations. Although texts on good research practice can include such broader considerations, appraisal tools tend to focus on technical quality and internal validity. A greater emphasis on perspectives could enable an increased clarity about the way that perspectives frame both the use and production of research and the effects of this on debates about the worth of competing evidence claims.

Types of Claim: Research Questions, Methods, and Paradigms

In addition to the perspectives driving research questions, research is very diverse in the types of questions that are asked and the methods used to address these. Education research alone has many different research themes and methods. In addition, there are further disciplines such as psychology and economics that study education topics with their own particular research focus and research designs.

Research can investigate many aspects of a phenomenon including its nature and extent, the impact of an intervention or service, and the processes by which something happens or is understood. Research can study many forms of evidence including experiential and tacit knowledge. It can create findings that describe, measure, compare, relate, and assess value. Within the wide diversity of research approaches, a common distinction is between inductive and deductive paradigms for relating theory to data created and interpreted within such theories.

Inductive approaches tend to ask open questions and use iterative configuring methods and analysis leading to results framed in terms of conceptual inference. Such research may change the way that people understand issues sometimes called the conceptual or enlightenment use of research (Weiss, 1979). Some of this research may be investigating how people perceive or experience some phenomena. Other research may be trying to develop new explanatory concepts or theories. The studies often use configuring methods of analysis of qualitative data.

Deductive approaches tend to ask closed questions and use a priori aggregative methods and analysis leading to test hypotheses framed in terms of instrumental inference. Such research provides "facts" that can lead to instrumental decision making (Weiss, 1979). Some of this research is based on statistics and the probability that something does or does not pertain. Some of these studies are based on sampling from specific groups and contexts and then generalizing to the rest of the population from which the sample was drawn. The studies often use aggregative methods of analysis of quantitative data.

Although it is common for researchers to gravitate to one or other of these paradigms, both make important contributions to developing and assessing theory and data (Oakley, 2000). Both make evidence claims related to theory and to empirical data at different levels of analysis.

Scope, Level, Focus, and Generalizability of Evidence Claims

Evidence claims also vary in their scope in terms of the breadth of the issue that they address, and the depth of detail of the claim (Gough et al., 2019). An evidence claim at one level may mask a claim at a more micro or macro level of analysis. A study of educational outcomes and the processes by which these outcomes occurred for all school students may, for example, mask important differences between subgroups of those students.

It may be that research is not asking the most relevant questions for potential users of that research. If the research questions are framed with a broader or narrower or more macro or more micro scope they might lead to different findings and implications. As discussed earlier, there is also the possibility of alternative explanations for research findings using very different perspectives and theoretical lenses. Critical appraisal includes the considerations of such alternative explanations. Are there alternative ways of framing the issues that might be more helpful or more parsimonious? Are there credible alternative explanations for this evidence having taken account of the strength of the evidence claim and the exploration and exclusion of these alternative explanations?

Some research is also undertaken to provide findings with generic knowledge while others make more context-specific claims (Oliver et al., 2018). If an evidence claim includes generalizations (e.g., in a theory, theory of change, or mechanisms), then it may have a wide scope relevant to many contexts. It could be that one piece of research evidence may be used to make strong evidence claims that are transferable to particular contexts and more limited claims for more generic contexts. There are tools to assess the extent of generalizability of such claims, in other words, the extent of the warrant (Burchett et al., 2020). Evidence claims may also refer to the implementation and scale-up of policy and practice decisions informed by research evidence.

With different perspectives producing and interpreting evidence according to different logics, it may not always be possible to adjudicate between the strength and veracity of different evidence claims. Two seemingly contradictory evidence claims may both be technically and logical defensible if they are produced from very different worldviews, research questions, and types of analysis. Users of research depend on transparency in the rationale for the evidence claim including any data, analysis, logic, and underlying perspectives to assess the claims being made. Their task might be made easier and research maybe more productive if there was also more explicit layering of the levels of research questions and the scope of related evidence claims.

The Basis of the Warrant for Making an Evidence Claim

Toulmin (2003) argued that an evidence claim is based on a warrant linking data to the claim plus any qualifications or exceptions to the claim. An evidence claim might not be justified either through any deficiencies in the data or in the way that the warrant linked it to the evidence claim (including a warrant extrapolating beyond what was justified by the data).

Different disciplines and research paradigms vary in their logic linking the data to the claim. They vary in their "fields of argument" resulting in "variability or field-dependence of the backing needed to establish our warrants" (Toulmin, 2003, p. 96). Appraising an evidence claim thus requires an assessment not only of the technical quality of the execution of a study but also of the relevance and coherence of the logic being applied (Toulmin, 2003).In addition, even if the execution and logic are technically appropriate and coherent there may be aspects of the way that these logics are applied in practice that reduce the relevance of the findings to the research question being asked. One example, would be the way that a variable was measured that made the data less relevant to the question being asked. This results in three components: (1) quality of execution of the study producing the data, (2) the appropriateness of the method applied, and (3) the relevance of focus of the study (Gough, 2007).

Certainty of an Evidence Claim and Decision Making

Evidence claims vary in what is claimed but also in the degree of certainty of the claim, which depends on the strength of the evidence including its extent. Certainty may also be expressed in terms of probabilistic statements. The degree of certainty

(and thus uncertainty) is a key part of the evidence claim (Sense About Science, 2013). The Educational Endowment Foundation, for example, provides a tool kit summarizing research evidence of the efficacy of different educational interventions and distinguishes the statistical size of the effect (in terms of average months progress for the school students) from the certainty that that effect size is correct (evidence strength).[2]

The specific meaning of certainty varies depending on the research questions and methods being applied. One of the reasons to undertake research is to reduce uncertainty, but uncertainty is always there.

For example, in studies evaluating the impact of an intervention, certainty can include confidence in statistical conclusions about effect sizes, causal claims, and generalizability. The certainty of a positive effect has also to be balanced with the possibility of a harm or negative effect, whether or not any such evidence is available. The lack of relevant evidence is not the same as evidence of no effect. Our knowledge is always provisional based on current knowledge and that may change with further theoretical and empirical research.

In many cases judgments about the certainty of the evidence available is determined on the basis of standard agreed criteria. A particular level of certainty is required in order for decisions (and resultant actions) to be supported by that evidence.

There may also be situations where the general criteria for certainty are not met yet the evidence is suggestive of a certain result, and this may inform decision making. If there is little cost or risk in taking a course of action, then there is little to be lost by following weak evidence and taking such action. So, the minimum level of certainty of evidence can depend upon attitudes to risk and the opportunity costs involved.

THE BASIS FOR THE EVIDENCE CLAIM

How Is Evidence Brought Together to Make the Evidence Claim

Making an evidence claim based on the warrant of what is known from research requires the consideration of all of the current evidence base relevant to that. Evidence claims based on limited evidence from a single or a few primary research studies may be very limited in the claims that can be made.

For a researcher the most pressing issue may be the trustworthiness of their own particular study. But the findings of single studies, however technically excellent, may differ from those of all the other technically good-quality relevant research studies. In other words, evidence claims from single studies may be a form of selection bias in that they do not represent all of the research evidence available. What is required is an evidence claim based on all of the current relevant evidence base.

The argument is that any evidence claim can be appraised according to the extent that it is appropriately considering all of the relevant evidence rather than any one study. Evaluating the methodological adequacy of an individual study is a building block of evaluating the whole evidence base. Appraising an individual study alone is not likely to provide a very useful evidence claim to inform decisions (except of course when there is only a single study available).

In this way of thinking, the evidence base (rather than individual studies) is a starting point for policy, practice, and individual decision makers. It is also the starting point for planning new research in terms of how it might change the preexisting evidence base. In, for example, undertaking a power calculation in an effectiveness study, the usual focus is the sample size necessary to reveal a statistical effect in the primary study. Taking an evidence-based approach, however, the calculation is on the sample size necessary to reveal a change in results from the existing systematic review on the topic (Elliott et al., 2017). This contrasts with the practice of some funders of research in not requiring a review of the evidence base before funding new research. This is akin to "going shopping without first seeing what you have in the kitchen cupboard."

Bringing together what is known from all research is itself a research exercise and the methods used are commonly called a "systematic review," which can be defined as "a review of existing research using explicit, accountable rigorous research methods" (Gough et al., 2017, p. 2). Reviews of research are similar to primary research but at a "meta level." Instead of collecting primary data, their samples are the findings from preexisting relevant primary studies. Instead of analyzing primary data, they synthesize the findings from such studies. There are also reviews of reviews where the data are the findings of the included reviews.

Some reviews of research literature are undertaken by topic specialists. Such expert reviews can benefit from the knowledge, skills, and experience of the author, but a lack of formal methods of review may lead to bias in sample selection and analysis (just as a lack of formal process might do in primary research). There may not be clarity about the authors' perspectives and expertise, the methods used to review the literature, and whether the expertise is deeper in some aspects of the topic reviewed than others.

The logic of using systematic research processes to review research evidence applies to all research questions and evidence claims. The methods of review tend to reflect the research paradigms and logic of the research methods of the primary studies addressing the same type of research questions. There is therefore a wide range of methods of systematic review methods including statistical meta-analysis of experimental data, theory-driven reviews of causal processes, meta-ethnography, and multicomponent mixed-methods reviews (Gough, Davies, et al., 2020; Gough et al., 2017; Gough et al., 2019).

Technical Quality and Relevance of an Evidence Base

Critical appraisal of the evidence claims of a review of the relevant research requires consideration of how the evidence was brought together, in other words, the methods of the review. If the review was not undertaken in a technically excellent and relevant way, then less trust can be put on the evidence claims that it produces. There is also a need to check the quality and relevance of the research studies included in the review and thus contributing to the evidence claim. Finally, there may be issues of quality in terms of the research evidence as a whole. Whatever the quality of the included studies it may be that the evidence they produce is not very useful. This

provides us with the following three components of critical appraisal of an evidence claim about an evidence base, which are discussed in turn:

1. *Review methods:* How was the evidence on which the claim is made brought together?
2. *Included studies:* What is their quality and relevance to the review question?
3. *The totality of evidence from the review:* What is the nature and extent of the evidence and its ability to answer the review question (and to justify the warrant of the conclusions (evidence claim)?

Review Methods Appraisal

The review is the method for bringing together evidence to make an evidence claim. The appraisal of the technical quality and relevance of such reviews can be divided into the three dimensions introduced earlier under the section "Technical Quality and Relevance of an Evidence Base" (Gough, 2007). In terms of reviews of an evidence base these dimensions are the following:

1. *Quality of execution of the review method (within the requirements of that method).* This might include, for example, the methods of clarifying the research question, the inclusion criteria, search strategy, screening, mapping of studies, quality appraisal of the evidence, and synthesis to answer the review question). If the review is not undertaken well then less trust can be put on its findings. This includes internal quality assurance measures to maintain and appraise the execution of the methods.
2. *The appropriateness of the review method (for the review question being asked).* A review may be undertaken in a technically excellent way, but that review method may not be the best method for making that evidence claim.
3. *Relevance of the focus of the review (of how the review method is applied).* A review method may be technically excellent in execution and appropriate, but the way in which it is operationalized in practice may not be so relevant. The ways of measuring the phenomena being studied or any outcome variables, for example, may not be considered valid or reliable or the review may not be up-to-date and include all current known evidence. Ethical issues of how the study was prioritized and undertaken may also be an issue.

The first dimension on execution of the review method is a generic form of appraisal for that method as it is independent of the specific research question being asked. In other words, has the chosen method being executed correctly? The other dimensions of the appropriateness of method and relevance of focus are "review specific" in that they depend upon the review question being asked. In other words, are the methods and the focus taken appropriate for addressing the research question?

Different fields of argument and their various methods of research will have different criteria for appraising these three dimensions. With inductive research there

may be particular concerns about the way that the research represents the nature and contexts of what has been studied. With deductive research there may be concerns about the presence of bias that might produce misleading research results.

Undertaking these appraisals is dependent on transparent reporting, which is part of the quality of execution of a review. Without details of the perspectives, research question, technical methods, and internal quality assurance of a review, it not possible to appraise its quality or its evidence claim. In practice, the lack of availability of such information might lead the reader to assume that the review was not undertaken well. It is not difficult to find reports with the title of systematic review yet that lack clear reporting of methods or have obvious methodological deficiencies such as omitting studies that meet the review's inclusion criteria.

Reviews of research, like all research, vary in how they balance the resources available with their scope, quality of execution, and fitness for the appropriateness of method and focus. As an example, "rapid reviews" are popular as being undertaken rapidly they may be able to provide quicker answers for users of research. Also, the shortness in time may also be reflected in lower costs for funders. Rapidity is often achieved through some restriction in the review process, such as a limited search strategy, narrow scope, less detailed data and analysis, or few internal quality assurance processes. The issue then is how the benefits of rapidity balance with any limitations in the justifiable evidence claims arising from restrictions in the methods of review (Gough et al., 2019).

Included Studies Appraisal

The quality and relevance of a review are also dependent on the quality and relevance of the studies included in a review. These studies can be appraised using the same three dimensions used to appraise a whole review (of quality of execution of the research method, the appropriateness of that method, and the focus of the studies). The dimensions of appropriateness and relevance, however, are specific to the review being undertaken (rather than the aims of the primary study authors) as they are judged on how well they fit the review question.

1. *Quality of execution* of the research methods of the included studies (within the requirements of that method)
2. *The appropriateness of the research methods* of the included studies (for answering the review question)
3. *Relevance of the focus* of the application of the research method of included studies (for answering the review question)

Totality of Evidence Appraisal

The evidence from all of the quality appraised included studies as a whole (not just the studies individually) affects the evidence claim in terms of the nature and extent of the evidence. The *nature of the evidence* will be a function of what has been studied and how it has been studied and the type of evidence produced. The *extent of*

the evidence will be determined by the number and size of these studies and the certainty of the findings that this produces. It may be that there is only one or few relevant studies and these cannot well represent the phenomena being studied. It may also be that there are many technically excellent and relevant studies but the evidence that they have produced is not that helpful (low strength of evidence) in answering the research questions being asked. There is also the issue of the extent that alternative perspectives and explanations for the research findings have been explored. These may undermine the evidence claim being made.

APPRAISAL OF EVIDENCE CLAIMS

Evidence Standards

Whatever the research question, the methods of synthesis review, included studies, the totality of evidence, and the consideration of the basis of the warrant in making an evidence claim, there must be some criteria for making such decisions.

Evidence standards is a term that can be used to describe such criteria. A complication is that the term *evidence standards* is used in a number of different overlapping ways. Examples include the following.

Methods Guidance

One approach is the specification of standards for selecting appropriate research methods and how to evaluate how well this has been achieved. One example is standards for undertaking program evaluations (Yarbrough et al., 2010). Another is *meta-evaluation* (Scriven, 1969), though this term also has multiple meanings (Gough et al., 2014).

Methods Standards

These are criteria for making justifiable evidence claims and thus the main focus of this chapter. The criteria provide standards for what needs to be achieved and thus also appraising whether this has been achieved. This could be for a whole evidence claim or some subpart such as the appraisal of the quality of individual studies making up a broader evidence claim (as described in Sections I and II of Table 1). The criteria will depend on the questions being made, the methods being used, and the claims being made. Appraising an evidence claim about the impact of an intervention is, for example, different from the appraisal of a conceptual model.

Internal Quality Assurance

Appraisals of the quality and relevance of research are based not only on the selection of overall methods. They are also dependent on details of how the method is executed. Internal quality assurance enables those details to be monitored, managed, and reported.

Criteria to Justify a Policy or Practice

This refers to the extent that a policy or practice meets a particular evidence standards criteria to justify its use. Initially a new initiative may have little research evidence to supports its benefits, but over time there may be some indication to suggest that it is promising or stronger evidence in support.

Reporting Standards

The appraisals of research methods and evidence claims are often distinguished from standards for how research should be reported. In practice, though, there is some overlap. Transparency is an essential component of research in providing accountability of how research findings are produced. It is not surprising then that advice on how to undertake and report research focuses on the quality and relevance of the methods used.

AERA has standards for reporting research that are explained as follows:

> First, reports of empirical research should be warranted; that is, adequate evidence should be provided to justify the results and conclusions. Second, reports of empirical research should be transparent; that is, reporting should make explicit the logic of inquiry and activities that led from the development of the initial interest, topic, problem, or research question; through the definition, collection, and analysis of data or empirical evidence; to the articulated outcomes of the study. Reporting that takes these principles into account permits scholars to understand one another's work, prepares that work for public scrutiny, and enables others to use that work. These standards are therefore intended to promote empirical research reporting that is warranted and transparent. (AERA, 2006, p. 33)

The AERA reporting standards do not, however, define the conduct of research or the format in which it is reported, nor does meeting of the standards necessarily mean that the research is adequate or appropriate (AERA, 2006, p. 33).

Another example of reporting standards is the CONSORT Statement[3] (Consolidated Standards of Reporting Trials) for reporting randomized controlled trials. It is a 25-item checklist about how the trial was designed, analyzed, and interpreted, plus a flow diagram describing the progress of the research participants through the trial. There are 17 extensions of the statement for particular types of designs interventions and data including the reporting of where equity issues are relevant to the research.

There could be benefits in terms of quality and consistency if there were more agreed standards for how research is undertaken, appraised, and presented. It might also help to show where the differences are in perspectives as much as in methods. The reporting guidance of AERA and the move toward agreed ways to report and assess different types of research methods such as CONSORT and GRADE (Grading of Recommendations Assessment, Development and Evaluation) could assist with this. On the other hand, there are dangers of fixed standards limiting plurality and methodological development. However, there are such basic problems in the application standards (e.g., discussed later of evidence claims on "one or two studies") that there at least grounds for agreed principles if not specific methods.

Evidence Standards Tools and Guides

Tools and guides have been developed to advise people in how to achieve these standards and to appraise whether they have been achieved. These resources enable all these judgments about evidence standards to be more transparent, systematic, and accountable, and open to debate.

Tools and guides vary in the work that they do in enabling evidence standards. Some are relatively simple tools in providing checklists of what is required of a research method and the associated evidence claim. Others provide more detailed criteria and guides and to allow more fine-tuned appraisals. Others may in addition, or instead, provide texts explaining the principles that underlie such judgments and maybe also the processes by which such principles can be enacted. Principles and processes may be offered by experienced academic or government-level organizations including comprehensive rich texts on how research should be undertaken and errors to avoid and that may include formal internal quality assurance measures. These bodies may then rely to different extents on their experience and reputation in applying these principles and processes rather than set evidence standards criteria.

Scoring of and Evidence for Evidence Standards

Some appraisal tools and reporting standards offer scoring systems to enable an overall measure of acceptability of the research. The dangers with scores is that they may result in very different areas of weakness with different levels of seriousness being given equal weight. Even if weighting adjustments are made, it may not be clear what the evidence is for that weighting. The scoring system may provide a spurious technical accuracy and thus undermine the potential benefit of tools offering accountability and debate of judgments.

For that reason, some appraisal tools such as GRADE (described later) assume a high level of credibility of research and then lower or increase that assumption based on structured judgments about the strengths and weaknesses of the methods used and commentary to explain those judgments. This allows the identification of the areas of concern and a justification for the decisions taken.

Another issue is the extent that guides, tools, and scoring systemas are themselves evidence informed. Studies and evidence claims are produced or appraised on the basis of threats to the evidence claims. Many of the criteria are developed from the logic of the research methods that they are based on. In some cases they are based on research evidence of the effects of different practices. In statistical research, for example, the effects of different practices on biasing results can be estimated. In more conceptual iterative inductive research there may be a logical basis for different research approaches and a concern that extensive standardization might limit the development of research practice.

Examples of Critical Appraisal Tools

This next section provides brief descriptions of a number of reporting standards and critical appraisal tools as a very partial list of examples of the type of tools available.

FIGURE 1
Main Focus of Appraisal of Evidence Claims of Different Appraisal Tools

SECTION II COMPONENTS	APPRAISAL TOOLS: ILLUSTRATIVE EXAMPLES	
1. Review methods i. Execution of review ii Appropriateness of method iii. Focus of method	AMSTAR EMMIE PRISMA RAMESES ROBIS	
2. Included studies i. Execution of study ii. Appropriateness of method iii. Focus	AERA CASP	CERQual GRADE IES WWC
3. Totality of evidence i. Nature of evidence ii. Extent of evidence		
iii. Alternative explanation		

Note. AMSTAR = A Measurement Tool to Assess Systematic Reviews; EMMIE = an evidence rating scale to Encourage Mixed-Method crime prevention synthesis reviews; PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RAMESES = Realist and Meta-Narrative Evidence Syntheses: Evolving Standards; ROBIS = Risk of Bias in Systematic Reviews; CERQual = Confidence in Evidence from Reviews of Qualitative research; GRADE = Grading of Recommendations Assessment, Development and Evaluation; IES WWC = Institute of Education Sciences What Works Clearinghouse; AERA = American Educational Research Association; CASP = Critical Appraisal Skills Programme.

Figure 1 lists the examples and indicates where they fit within the technical quality and relevance part of Section II of the framework described in this chapter (and summarized in Table 1). These are discussed briefly below and described further in the Online Appendix (available in the online version of the journal).

Review Methods (AMSTAR 2, EMMIE, PRISMA, RAMESES, ROBIS)

Some tools focus on the reporting or appraisal of the review process that produces the findings rather than of the findings themselves. They tend to list the main components that should be included in a review with some explanation of the methodological logic behind this. As different review methods are used for different types of review questions, their methodological focus and what is considered important to include in a review differ.

Some reviews are concerned with hypothesis testing of effectiveness questions using inductive a priori paradigms. PRISMA[4] (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) provides guidance for the reporting of the main components of such effectiveness reviews whilst AMSTAR 2[5] (A Measurement Tool to Assess Systematic Reviews) is for the appraisal of such completed reviews (Shea et al., 2017).

ROBIS[6] (Risk of Bias in Systematic Reviews) is a tool for appraising the risk of bias in systematic reviews of effectiveness, etiology, and diagnostic test accuracy in health care (Whiting et al., 2016). The focus is on the trustworthiness of the review process, the individual studies contributing to the review, and the totality of the evidence found in the review. There are a number of studies comparing the nature and use of ROBIS and AMSTAR such as by Gates et al. (2020).

EMMIE (an evidence rating scale to Encourage Mixed-Method crime prevention synthesis reviews) is for appraising and planning reviews of effectiveness in crime prevention (Johnson et al., 2015). It is concerned with theory and the mechanisms by which interventions have their effect and so includes moderators and mediators as well as overall effects of interventions. Knowing how something might work is important to users of evidence in the transferability of interventions to new contexts.

RAMESES[7] (Realist and Meta-Narrative Evidence Syntheses: Evolving Standards) provides reporting guidance for realist reviews. These are also theory-driven effectiveness reviews of "what works for whom under what circumstances, how and why?" Realist reviews differ from some theory driven reviews in taking an iterative approach to the testing of theory: "iterative testing and refinement of theoretically based explanations using empirical findings in data sources" (Wong et al., 2013, p. 10). They have a first stage of configuring theoretical explanations and an inductive iterative stage for empirically testing the theory (Gough, 2013).

Reporting standards are also available from many other sources including MECIR[8] (Methodological Expectations of Cochrane Intervention Reviews) and MECCIR[9] (Methodological Expectations of Campbell Collaboration Intervention Reviews) for expectations of Cochrane and Campbell Collaboration systematic reviews. There are of course also textbooks and papers providing guidance on reviews on the nature of systematic reviews and how they vary with the research question asked (Gough et al., 2017; Gough et al., 2019).

Included Studies and Resultant Evidence Base (CERQual, GRADE, IES WWC)

Another type of critical appraisal focuses on the quality of the studies included in a review and the totality of evidence that these produce (rather than the critical appraisal of the overall method of review).

GRADE[10] is such a tool that checks for statistical bias in effectiveness studies in health research (Guyatt et al., 2008). It rates the certainty that a research finding across studies is correct for each outcome variable on the basis of the research design

and then increases or decreases this rating dependent on more detailed examination of the data (and it also considers strength of recommendations informed by that evidence).

CERQual[11] (Confidence in Evidence From Reviews of Qualitative research) takes a similar approach to GRADE but is for checking the confidence in the results of reviews of qualitative research (Lewin et al., 2018). Instead of being concerned with statistical bias, the tool focuses on other forms of methodological limitations and the coherence, richness, and relevance of the data from primary studies for the review findings.

A third example is the U.S. Department of Education's Institute of Education Sciences What Works Clearinghouse (WWC).[12] This produces Evidence Reports summarizing the evidence base on the effectiveness of many educational interventions. It also produces practice guides on such interventions (see the Section "Engagement With Evidence: Communication, Fitness for Purpose, and Recommendations From Evidence").

Primary Research Study Guides and Tools (AERA, CASP Qualitative)

Historically most guides and tools were for appraising the technical quality and internal validity of the methods of primary research, and this learning feeds through into the appraisal of included studies in reviews of an evidence base. It has also been important in developing the conduct and reporting of primary research as in the already mentioned CONSORT system for reporting of quantitative research.

The CASP (Critical Appraisal Skills Programme) tool for qualitative research[13] is an example of a primary research appraisal tool. It is a short checklist and is one of a number of similar CASP tools for appraising different types of primary research and systematic reviews. CERQual also suggests its use in its appraisal system for qualitative research.

Another example in the Online Appendix is the AERA (2006) Standards for Research Conduct. These reporting standards take a broad approach in covering both quantitative and qualitative studies and also how research questions are framed. It is unusual in covering both the nature of claims as well as their technical quality and relevance (both I and II in Table 1).

Narrowness of Evidence Standards and Critical Appraisal Tools and Guides

Figure 1 shows how each of the tools consider particular components of the framework presented in Table 1. Many tools focus only on the technical quality of the methods of review, or of the included studies alone, or of included studies and the totality of evidence. In practice, the appraisal of included studies is a very common form of tool as these were developed prior to and then incorporated into a stage of

systematic review. Appraisal of review methods and the totality of evidence is more recent. All of the tools are concerned with the warrants for making an evidence claim. Sometimes this also includes the examination or control of alternative explanations for the findings on which the evidence claim is made though this is not a routine or always explicit part of warrant making.

Another form of narrowness of focus is that many tools are concerned with very specific research questions and specific research methods. This is not surprising as different research questions and associated methods raise different issues of quality. Aggregative research that examines questions of, for example, the evidence for the effectiveness of an intervention tends to be concerned about issues of bias from uncontrolled variables. Configurative research that is developing new conceptual understandings may be more concerned with issues of the richness or veracity with which the research reflects the object of study.

This focus on specific research questions also means that the tools examine a relatively homogeneous type of studies. Although there is increased use of mixed-methods research, this is often multicomponent research with appraisal methods varying across the components.

ENGAGEMENT WITH EVIDENCE

Communication of Evidence Through Web-Based Evidence Portals

Journals, books, and the internet provide many opportunities for evidence claims and recommendations derived from such claims to be made widely available. This can be a strength and a weakness. The number of studies can be overwhelming. The easy availability of studies in many forms may be challenging to understand, particularly when written in technical language. Summaries and reviews of research evidence may provide easy access, but the extent of any formal process of review may limit the confidence that can be made about the evidence claim conclusions.

One way in which research evidence is made available to enable the use of that research in policy and practice decision making is the provision of curated summaries of research findings in evidence portals. Another method is to provide practice recommendations or guidance based on that evidence.

For research evidence to inform decision making then it needs to be accessible to policy, practice, and individual decision makers (though other factors are of course also necessary; Langer et al., 2016). For this reason, some knowledge intermediary (brokerage) organizations provide a service by providing summaries of research evidence in consumer guides, web-based evidence portals, research clearinghouses, and evidence tool kits. It is more efficient for such organizations to assemble and communicate this evidence than for all individual decision makers to do so independently. This, however, raises the issue of the evidence standards used by the organizations to communicate this research evidence (Gough et al., 2018; Gough, Maidment, & Sharples, 2020; Means et al., 2015).

TABLE 2
Web-Based Portals' Evidence Standards for Making Evidence Claims

Basis for Evidence Claim	Evidence Tool Kits
Systematic reviews	6
Narrative, expert reviews or listing of studies	4
1 or 2 "good studies"	5
Total	15

Source. Adapted from Gough and White (2018).

Table 2 summarizes the results of a simple survey of 15 such evidence portals (Gough & White, 2018). This survey found that only 6 of the 15 portals used systematic reviews to make evidence claims. Five of them made evidence claims about whether an intervention had evidence of being effective on the basis of two good-quality studies showing a positive effect (and for some, no evidence of harmful effects). This is obviously a weak basis for an evidence claim as, however good the one or two studies are, this is unlikely to include all of the relevant evidence base. There may be many other well conducted relevant studies that did not find this positive effect. Basing evidence claims on the basis of one or two studies is a lower standard than "vote counting," where the number of studies showing a positive effect is compared with the number of studies showing no evidence of any effect or a negative effect. Such low standards for making evidence claims can undermine the role of knowledge brokerage in promoting the use of research in informing decision making.

One of the tool kits with the evidence claim criteria of "1 or 2 good studies" included in the 2018 survey was the U.S. WWC in education. However, the criteria was increased in January 2020 (from WWC Procedures Handbook 4.0 to 4.1) so that strength of evidence is assessed using a systematic review of the relevant evidence base using fixed-effects meta-analysis.[11] At the time of writing, the U.S. Department of Education's Every Student Succeeds Act still does not require a systematic synthesis for Tier 1 Strong Evidence. Strong Evidence is defined as at least one well-designed and well-implemented experimental study (with a large multisite sample that overlaps with the populations and settings proposed to receive the intervention) though it also requires that the findings are not overridden by statistically significant and negative evidence on the same intervention in other studies (Department of Education, 2016, p. 8).

Fitness for Purpose of Evidence Claims

This chapter has so far considered the components of appraising whether an evidence claim is justifiable. If the evidence is not trustworthy, it cannot be relied on. This is a necessary but not sufficient step for a decision maker using evidence. The decision maker also needs to ensure that the evidence is fit for purpose in other ways.

Research evidence can inform and researchers can advise but cannot alone make the decision. Mention has already been made about how research may vary in the perspectives driving how it addresses issues, its availability, its certainty, and its consistency across a range of issues involved in a decision. Policy, practice, and individual decision making is likely to include the consideration of a wide range of perspectives and types of information beyond research, and is inherently a values-based process. Providing evidence to inform a decision is thus very different from making a policy, practice, or individual decision, which needs to balance advantages and disadvantages of different courses of action.

Research is also undertaken in particular situations and may vary in its generalizability. Research varies in the extent that it is context dependent, in the extent that it is explicit about such generalizability, and in the extent that it provides information that enables transferability of research findings to new contexts (e.g., information on mechanisms, theories of change, and implementation of policies and practices). Research may also mislead if users are not aware of highly relevant alternative evidence claims that would be identified by different research questions or similar questions with a different scope.

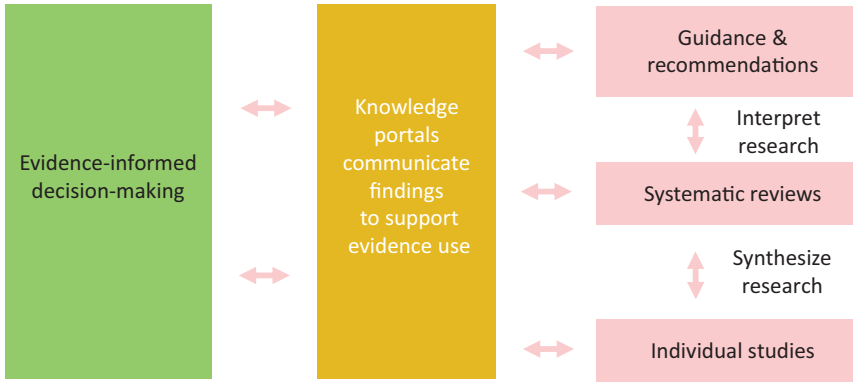
Research may also not provide the information for users to make such fit for purpose assessments. Systematic reviews, for example, may provide detailed descriptions and meet the specific methodological and reporting criteria required by appraisal tools such as AMSTAR and PRISMA, yet they may still not be fully transparent about, for example, the breadth and depth of the research question used and the consequences of such choices.

A review may be very explicit about what type of studies were included but not consider the implications of what was not included. A review that has a high bar of only including particular research designs is, in one way, being very rigorous, but it may misrepresent what is known if studies of other designs that can contribute meaningfully to the evidence base are excluded. Similarly, a more broadly framed review question might include studies that change the understanding of a narrowly framed review. The name "on the tin" of a the review and detailed reporting of rigorous review methods may not be sufficient to check the warrant for the evidence claim, including consideration of alternative explanations for the research findings. In sum, the way that evidence is framed (e.g., level of analysis, breadth of scope, framing of issues) may distort the evidence even if on a technical basis the evidence claim is strong. Finally, decisions are made on the basis of many factors, not just research evidence. There may also be many aspects to a decision informed by very different types of evidence with different levels of certainty. Decision making can therefore be a very complex process.

Recommendations, Guidance, and Practice Guidelines

The section above on fitness for purpose discussed how users of research evidence need to interpret the meaning and relevance of this evidence in terms of their particular contexts and other information available to them. Much of these contexts may be shared, particularly for professional practitioners, and so it might be more efficient

FIGURE 2
The Position of Evidence Portals in the Evidence Ecosystem



Source. Adapted from Gough and White (2018).

for there to be some form of local or national process for the interpretation and recommendations or guidance arising from the evidence.

Just as the synthesis of research findings is more efficient being undertaken using formal processes by specialist agencies rather than individual decision makers, the same argument can be applied to the production of recommendations for policy and practice. Knowledge portals can then instead of reporting on individual studies or reviews of such studies can communicate practice recommendation and guidance (see Figure 2).

Not everyone may agree on the particular interpretations made (and the range of perspectives driving these), but these perspectives and other information informing the decisions (including the research evidence used) can be the product of transparent systematic processes driven by explicit stakeholder representatives with the interpretations open to challenge and debate. Such recommendations and guidance may not have to be followed, but they provide a more efficient starting point than these activities being duplicated by many in society.

Just as there is a formal systematic approach to reviewing an evidence base, structured approaches have also been developed for making recommendations for policy and practice. GRADE, for example, uses an evidence to decision framework that examines the evidence on: the priority of the problem, the benefits and harms of what is recommended, the resources required to implement, the equity impacts, and the acceptability of the recommendations to stakeholders.

An example of a broader approach to guidance is the health and social care guidelines of the National Institute for Health and Care Guidance (NICE) a government-funded agency in the United Kingdom. The topics for guidance are first suggested by

the government Department of Health and Social Care. NICE then undertakes a stakeholder consultation exercise by email and through a meeting to help determine the focus of the review and to appoint a guidance committee that will manage the production of the guidance. This is stakeholder-driven guidance as the committee focuses the research questions to be addressed, commissions systematic reviews of the evidence, and interprets their findings in the light of the U.K. context and practice experience where there is a lack of research.

A similar approach is taken by the already mentioned U.S. WWC for education that provides Practice Guides in addition to intervention reports on the evidence about particular interventions. A Practice Guide is a publication that presents recommendations for educators to address challenges in their classrooms and schools. They are based on reviews of research, the experiences of practitioners, and the expert opinions of a panel of nationally recognized experts.

For a Practice Guide recommendation to be supported by a "strong evidence base" requires the following: (1) there is a preponderance of "positive effects" without contradictory of negative or potentially negative effects, (2) there is a "medium to large" extent of evidence, (3) the research has direct relevance to the scope of the practice, (4) the practice is a major component of the interventions tested in the studies, (5) the panel creating the guide has a high degree of confidence that the practice is effective, and (6) when assessment is the focus of the recommendation they meet the Standards for Educational and Psychological testing. The criteria for what is a strong evidence base were strengthened in January 2020 although many existing practice guides may have been produced before this date.

Both NICE and the WWC develop guidance using systematic review methods supported with detailed handbooks on research methods. NICE's topics are driven by government requirements for guidance in topic areas and are then focused by a user stakeholder committee. The WWC panels represent academic and practitioner specialists.[14] The NICE stakeholders also include users of the services.

For health practice guidelines there are also tools to appraise the quality and relevance of guidelines. For example, the AGREE[15] instrument addresses the methodological quality and reporting of guidelines while AGREE REX examines their clinical credibility and implementability. A further issue is how research guidance is used in practice. Guidance may provide details of how this should be done and what skills, contexts, or further information are required to do this. Such use of guidance can also be the subject of research study. This is one part of the broader topic of how evidence claims are used in practice and "research on research use."

CONCLUSION

This chapter has examined four different components for appraising the fitness for purpose of evidence claims at a high level of abstraction. The components are listed in Table 1 and the issues raised can be summarized as follows:

I. The evidence claim: *What is the claim and is it relevant to the users' needs?*
II. The basis for the evidence claim: *Is the claim warranted? Do other claims (maybe framed differently) need to be considered too?*
III. Appraisal of evidence claims, evidence standards, tools and guides: *Are the evidence standards used appropriate for making these warrants?*
IV. Engagement with evidence: *How is other information being used to interpret the evidence to inform decision making? Is the certainty provided by the evidence claim sufficient in absolute terms and in terms of the other information being considered and the opportunity costs of the decision being made?*

There are many much more detailed aspects of each of the components, which are too detailed for consideration here. The aim of the chapter has been to offer a high level overview of key issues in the fitness for purpose of evidence in terms of both its technical quality (and the veracity of evidence claims) and its fitness for use by policy, practice, and individual decision makers. It is in a sense locating fitness for purpose of evidence within an evidence ecosystem rather than as disconnected components of research use and research production.

ORCID iD

David Gough https://orcid.org/0000-0003-2732-0402

NOTES

[1]https://www.jla.nihr.ac.uk/

[2]https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/ (accessed March 11, 2020).

[3]http://www.consort-statement.org/ (accessed February 22, 2020).

[4]http://www.prisma-statement.org/ (accessed February 19, 2020).

[5]https://amstar.ca/index.php (accessed February 19, 2020).

[6]https://www.bristol.ac.uk/population-health-sciences/projects/robis/ (accessed February 19, 2020).

[7]https://www.ramesesproject.org/ (accessed March 11, 2020).

[8]https://methods.cochrane.org/methodological-expectations-cochrane-intervention-reviews (accessed July 5, 2020).

[9]https://campbellcollaboration.org/about-meccir.html (accessed July 5, 2020).

[10]https://gdt.gradepro.org/app/handbook/handbook.html (accessed February 20, 2020).

[11]https://www.cerqual.org/ (accessed March 13, 2020).

[12]https://ies.ed.gov/ncee/wwc/handbooks (accessed February 23, 2020).

[13]https://casp-uk.net/wp-content/uploads/2018/01/CASP-Qualitative-Checklist-2018.pdf (accessed February 20, 2020).

[14]"Practice guide topics are selected based on their potential to improve key outcomes, their applicability to a broad range of students or to particularly important subpopulations, their policy relevance, the perceived demand within the education community, and the availability of rigorous research to support recommendations" (WWC Procedures Handbook 4.1, p. A-3).

[15]https://www.agreetrust.org/ (accessed July 24, 2020).

REFERENCES

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*(6), 33–40. https://www.aera.net/Portals/38/docs/12ERv35n6_Standard4Report%20.pdf

American Educational Research Association. (2011). Code of ethics, American Educational Research Association, approved by the AERA council, February 2011. *Educational Researcher*, *40*(3), 145–156. https://cdn.ymaws.com/www.weraonline.org/resource/resmgr/a_general/aera.pdf

Best, A., & Holmes, B. (2010). Systems thinking, knowledge and action: Towards better 1227models and methods. *Evidence & Policy*, *6*(2), 145–159. https://doi.org/10.1332/174426410X502284

Burchett, H. E. D., Kneale, D., Blanchard, L., & Thomas, J. (2020). When assessing generalisability, focusing on differences in population or setting alone is insufficient. *Trials*, *21*(1), Article 286. https://doi.org/10.1186/s13063-020-4178-6

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *Lancet*, *374*(9683), 86–89. https://doi.org/10.1016/S0140-6736(09)60329-9

Colvin, C. J, Garside, R., Wainwright, M., Lewin, S., Bohren, M., Glenton, C., Munthe-Kaas, H. M., Carlsen, B., Tuncalp, Ö., Noyes, J., Booth, A., Rashidian, A., Flottorp, S., & Lewin, S. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 4: how to assess coherence. *Implementation Science*, *13*(Suppl. 1), Article 13. https://doi.org/10.1186/s13012-017-0691-8

Department of Education. (2016). *Non-regulatory guidance: Using evidence to strengthen education investments*. https://www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

Djulbegovic, B., & Guyatt, G. H. (2017). Progress in evidence-based medicine: A quarter century on. *Lancet*, *390*(10092), 415–423. https://doi.org/10.1016/S0140-6736(16)31592-6

Elliott, J., Synnot, A., Turner, T., Simmonds, M., Akl, E., McDonald, S., Salanti, G., Thomas, J., Meerpohl, J., MacLehose, H., Hilton, J., Shemilt, I., Tovey, D., & Living Systematic Review Network. (2017). Living systematic review: 1. Introduction-the why, what, when, and how. *Journal of Clinical Epidemiology*, *91*, 23–30. . https://doi.org/10.1016/j.jclinepi.2017.08.010

Gates, M., Gates, A., Duarte, G., Cary, M., Becker, M., Prediger, B., Vandermeer, B., Fernandes, R. M., Pieper, D., & Hartling, L. (2020). Quality and risk of bias appraisals of systematic reviews are inconsistent across reviewers and centers. *Journal of Clinical Epidemiology*, *125*, 9–15. https://doi.org/10.1016/j.jclinepi.2020.04.026

Glenton, C., Carlsen, B., Lewin, S., Munthe-Kaas, H. M., Colvin, C. J., Tuncalp, Ö., Bohren, M., Noyes, J., Booth, A., Garside, R., Rashidian, A., Flottorp, S., & Wainwright, M. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 5: How to assess adequacy of data. *Implementation Science*, *13*(Suppl. 1), 14. https://doi.org/10.1186/s13012-017-0692-7

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. https://doi.org/10.1080/02671520701296189

Gough, D. (2013). Meta-narrative and realist reviews: Guidance, rules, publication standards and quality appraisal. *BMC Medicine*, *11*, Article 22. https://doi.org/10.1186/1741-7015-11-22

Gough, D., Davies, P., Jamtvedt, G., Langlois, E., Littell, J., Lotfi, T., Masset, E., Merlin, T., Pullin, A., Ritskes-Hoitinga, M., Røttingen, J-A., Sena, E., Stewart, R., Tovey, D., White, H., Yost, J., Lund, H., & Grimshaw, J. (2020). Evidence synthesis international position statement. *Systematic Reviews*, *9*, Article 155. https://doi.org/10.1186/s13643-020-01415-5

Gough, D., & Elbourne, D. (2002). Systematic research synthesis to inform policy, practice and democratic debate. *Social Policy and Society*, *1*(3), 1–12. https://doi.org/10.1017/S147474640200307X

Gough, D., Maidment, C., & Sharples, J. (2018). *UK What Works Centres: Aims, methods and contexts*. EPPI-Centre, Social Science Research Unit, UCL 1233, Institute of Education, University College London.

Gough, D., Maidment, C., & Sharples, J. (2020). *Enabling evidence-informed policy and practice to be evidence informed*. Manuscript submitted for publication.

Gough, D., Martin, S., Bovaird, T., & France, J. (2014). *Meta-evaluation of the impact and legacy of the London 2012 Olympic Games and Paralympic Games: ESRC methods paper*. EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London.

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews* (2nd ed.). Sage.

Gough, D., Thomas, J., & Oliver, S. (2019). Clarifying differences between reviews within evidence ecosystems. *Systematic Reviews*, *8*(1), Article 170. https://doi.org/10.1186/s13643-019-1089-2

Gough, D., & White, H. (2018). *Evidence standards and evidence claims in web based research portals*. Centre for Homelessness Impact.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J., & GRADE Working Group. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Johnson, S., Tilley, N., & Bowers, K. J. (2015). Introducing EMMIE: An evidence rating scale to encourage mixed-method crime prevention synthesis reviews. *Journal of Experimental Criminology*, *11*, 459–473. https://doi.org/10.1007/s11292-015-9238-7

Langer, L., Tripney, J., & Gough, D. (2016). *The science of using science: Researching the use of research evidence in decision-making*. EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London.

Lewin, S., Bohren, M., Rashidian, A., Glenton, C., Munthe-Kaas, H. M., Carlsen, B., Colvin, C. J., Tuncalp, Ö., Noyes, J., Booth, A., Tunçalp, Ö., Wainwright, M., Flottorp, S., Tucker, J. D., & Carlsen, B. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 2: How to make an overall CERQual assessment of confidence and create a Summary of Qualitative Findings table. *Implementation Science*, *13*(Suppl. 1), 10. https://doi.org/10.1186/s13012-017-0689-2

Means, S. N., Magura, S., Burkhardt, J. T., Schröter, D. C., & Coryn, C. L. S. (2015). Comparing rating paradigms for evidence-based program registers in behavioral health: Evidentiary criteria and implications for assessing programs. *Evaluation and Program Planning*, *48*, 100–116. https://doi.org/10.1016/j.evalprogplan.2014.09.007

Munthe-Kaas, H. M., Bohren, M., Carlsen, B., Glenton, C., Lewin, S., Colvin, C. J., Tuncalp, Ö., Noyes, J., Booth, A., Garside, R., Colvin, C. J., Wainwright, M., Rashidian, A., Flottorp, S., & Carlsen, B. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 3: How to assess methodological limitations. *Implementation Science*, *13*(Suppl 1), 9. https://doi.org/10.1186/s13012-017-0690-9

Noyes, J., Booth, A., Lewin, S., Carlsen, B., Glenton, C., Munthe-Kaas, H. M., Colvin, C. J., Garside, R., Bohren, M., Rashidian, A., Wainwright, M., Tunçalp, Ö., Chandler, J., Flottorp, S., Pantoja, T., Tucker, J. D., & Munthe-Kaas, H. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 6: How to assess relevance of the data. *Implementation Science*, *13*(Suppl. 1), 4. https://doi.org/10.1186/s13012-017-0693-6.

Oakley, A. (2000). *Gender and method in the social sciences*. Polity Press.

Oliver, S., Roche, C., Stewart, R., Bangpan, M., Dickson, K., Pells, K., Cartwright, N., Hargreaves, J., & Gough, D. (2018). *Stakeholder engagement for development impact evaluation and evidence synthesis* (CEDIL Inception Paper 3). Centre of Excellence for Development Impact and Learning. https://cedilprogramme.org/publications/stakeholder-engagement-for-development-impact-evaluation-and-evidence-synthesis/

Scriven, M. (1969). An Introduction to meta-evaluation. *Educational Products Report*, *2*(5), 36–38.

Sense About Science. (2013). *Making sense of uncertainty*. https://senseaboutscience.org/wp-content/uploads/2016/11/Makingsenseofuncertainty.pdf

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*, *7*, 358. https://doi.org/10.1136/bmj.j4008

Stenhouse, L. (1981). What counts as research. *British Journal of Educational Studies*, *29*(2), 103–114. https://doi.org/10.1080/00071005.1981.9973589

Toulmin, S. (2003). *The uses of argument*. Cambridge University Press.

Weiss, C. (1979). The many meanings of research utilization. *Public Administration Review*, *39*(5), 426–431. https://doi.org/10.2307/3109916

Welch, V., Jull, J., Petkovic, J., Armstrong, R., Boyer, Y., Cuervo, L. G., Edwards, S. J. L., Lydiatt, A., Gough, D., Grimshaw, J., Kristjansson, E., Mbuagbaw, L., Mcgowan, J., Moher, D., Pantoja, T., Petticrew, M., Pottie, K., Räder, T., Shea, B., . . . Tugwell, P. (2015). Protocol for the development of a CONSORT-equity guideline to improve reporting of health equity in randomized trials. *Implementation Science*, *10*, 146. https://doi.org/10.1186/s13012-015-0332-z

What Works Clearinghouse. (2020). *What Works Clearinghouse procedures handbook, Version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/handbooks

What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook, Version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/handbooks

Whiting, P., Savovic, J., Higgins, J. P., Caldwell, D. M., Reeves, B. C., Shea, B., Davies, P., Kleijnen, J., Churchill, R., & the ROBIS Group. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology*, *69*, 225–234. https://doi.org/10.1016/j.jclinepi.2015.06.005

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., & Pawson, R. (2013). RAMESES publication standards: Realist syntheses. *BMC Medicine*, *11*, Article 21. https://doi.org/10.1186/1741-7015-11-21

Yarbrough, D. B., Shula, L. M., Hopson, R. K., & Caruthers, F. A. (2010). *The Program Evaluation Standards: A guide for evaluators and evaluation users* (3rd. ed.). Corwin Press.