PRE PUBLICATION COPY OF PAPER PUBLISHED IN:

Gough D (2007) Weight of evidence: a framework for the appraisal of the quality and relevance of evidence In J. Furlong, A. Oancea (Eds.) Applied and Practice-based Research. Special Edition of *Research Papers in Education*, 22, (2), 213-228

WEIGHT OF EVIDENCE : A FRAMEWORK FOR THE APPRAISAL OF THE QUALITY AND RELEVANCE OF EVIDENCE

David Gough

ABSTRACT

Knowledge use and production is complex and so also are attempts to judge its quality. Research synthesis is a set of formal processes to determine what is known from research in relation to different research questions and this process requires judgements of the quality and relevance of the research evidence considered. Such judgement can be according to generic standards or be specific to the review question. The judgements interact with other judgements in the review process such as inclusion criteria and search strategies and can be absolute or weighted judgements combined in a weight of evidence framework. Judgments also vary depending upon the type of review that can range from statistical meta analysis to meta ethnography. Empirical study of the ways that quality and relevance judgements are made can illuminate the nature of such decisions and their impact on epistemic and other domains of knowledge. Greater clarity about such ideological and theoretical differences can enable greater participative debates about such differences.

INTRODUCTION

Oancea and Furlong (this volume) suggest that there are a number of different domains that need to be considered when assessing quality in applied and practice based research; these domains they describe as the epistemic, the phronetic, and the technical and economic. For them, quality in applied and practice based research needs to be conceptualised more broadly than has conventionally been the case.

If research is to be of value in applied contexts, then these issues of quality cannot be judged only according to abstract generic criteria but must also include notions of fitness for purpose and relevance of research in answering different conceptual or empirical questions. In other words, question specific quality and relevance criteria are used to determine how much 'weight of evidence' should be given to the findings of a research study in answering a particular research question.

This paper addresses these issues with reference to systematic reviews and systematic research synthesis where a number of studies are considered individually to see how they then collectively can answer a research question. The paper is principally concerned with the quality and relevance appraisal of this epistemic knowledge. Providing greater clarity on how epistemic knowledge is developed and used can

make its role more transparent in relation to the other domains of knowledge described by Oancea and Furlong.

Systematic synthesis is a set of formal processes for bringing together different types of evidence so that we can be clear about what we know from research and how we know it (Gough and Elbourne 2002, Gough 2004). These processes include making judgements about the quality and relevance assessment of that evidence. The paper focuses on the systematic methods of research synthesis but systematic methods of synthesis and arguments of weight of evidence can be applied to the (epistemic) evaluation of all types of knowledge.

Being specific about what we know and how we know it requires us to become clearer about the nature of the evaluative judgements we are making about the questions that we are asking, the evidence we select, and the manner in which we appraise and use it. This then can contribute to our theoretical and empirical understanding of quality and relevance assessment. The questions that we ask of research are many and come from a variety of different individual and group perspectives with differing ideological and theoretical assumptions. In essence, the appraisal of evidence is an issue of examining and making explicit the plurality of what we know and can know.

The paper first sets the scene with a brief sketch of how research is just one of many activities concerned with knowledge production and its appraisal and use. Second, the paper introduces evidence synthesis and the crucial role of quality and relevance assessment in that process to judge how much 'weight' should be given to the findings of a research study in answering a review question. Finally, it is argued that we should study how judgements of quality are made in practice and thus develop our sophistication in quality appraisal and synthesis of research and other evidence.

ACTION, RESEARCH, KNOWLEDGE AND QUALITY APPRAISAL

We all act on and in the world in different ways and in doing so create different types of knowledge. The knowledge produced may be relatively tacit or explicit, it can be used to develop ways of understanding or more directly to inform action with varying effects, and it can produce 'capacity for use' or more direct technological value and economic and other impacts (Furlong and Oancea 2006). Particular groups of people tend to focus on particular activities and produce particular types of products. So researchers, for example, undertake research to produce knowledge and understanding and in doing so they probably also produce many other sorts of knowledge. Working as a researcher can provide experiences ranging from team working with colleagues and participants of research to the use of computer software and lead to organisational and practice knowledge about research (Pawson *et. al.* 2003). All these different types of knowledge can be used in different ways leading to different intended and unintended and direct and indirect effects. When there is overt use of knowledge, this use may include an appraisal of its fitness for purpose.

Table 1 Examples of knowledge production and use across different ideological and theoretical standpoints[1]

ACTORS	KNOWLEDGE TYPES	KNOWLEDGE EXPLICITNESS	KNOWLEDGE APPRAISAL	KNOWLEDGE USE	KNOWLEDGE IMPACT
Researcher	Research	Tacit to declarative dimension		Understanding to action dimension	Physical
Service user	Use				Social
Practitioner	Practice				Economic
Policy maker	Policy				
Organisational	Organisational				

Table 1 lists some of these main dimensions of the flow between action and knowledge production. These can be complex and interactive processes that involve different psychological and social mechanisms and rely on varying ideological and theoretical stand points. These different ideologies and theories may be mutually exclusive or even premised on the need actively to critique the assumptions and understandings of other perspectives.

The quality and relevance of all this knowledge can be based on generic criteria or in relation to some specific criteria and purpose. In relation to generic criteria, any object might be thought of as high quality because of the materials being used, the manner in which they have been put together, the beauty of the resulting object, its fitness for purpose or how form and function combine. For research knowledge, the research design and its execution is often considered important.

Use specific criteria may be even more varied. The processes of knowledge creation and use listed in Table 1 can be so complex and based on so many different theories and assumptions that it is difficult to independently determine what the use specific criteria should be for assessing quality and relevance of that knowledge. For example, a policy maker may have different assumptions about and criteria for evaluating policy, organisational and research knowledge and may apply knowledge developed and interpreted within these world views to achieve different physical, social and economic impacts. They may also use research knowledge to evaluate between policy choices or to support choices already made (Weiss 1979).

This complexity provides the background for the focus of this paper which is the quality and relevance appraisal of research knowledge. The concern is with the evaluation of studies in the context of research synthesis that considers all the research addressing the research questions being asked. Users of research (ranging from policy makers to service providers to members of the public) often want to ask what we know from all research as a whole rather than just considering one individual study.

EVIDENCE SYNTHESIS

Much of our use of knowledge is to answer such questions as ‘how do we conceptualise and understand this?’ or ‘what do we know about that?’. We can use what we know from different sorts of knowledge collected and interpreted in different

[1] Informed by Pawson *et. al.* (2003); Furlong and Oancea (2006)

ways to develop theories, test theories, and make statements about (socially constructed) facts.

So how do we bring together these different types of knowledge? Just as there are many methods of primary research there are a myriad of methods for synthesizing research which have different implications for quality and relevance criteria. A plurality of perspectives and approaches can be a strength if it is a result of many differing views contributing to a creative discussion of what we know and how we know it and what we could know and how we could know it. The challenge is to develop a language to represent this plurality to enable debate at the level of synthesis of knowledge rather than at the level of individual studies.

Systematic evidence synthesis reviews

Before discussing these approaches to reviewing literature, it may be helpful to clarify two confusing aspects of terminology about research reviews. The first issue is the use of the term 'systematic'. With both primary qualitative and quantitative research there is a common expectation that the research is undertaken with rigour according to some explicit method and with purpose, method and results being clearly described. All research is in a sense biased by its assumptions and methods but research using explicit rigorous methods is attempting to minimize bias and make hidden bias explicit and thus provide a basis for assessing the quality and relevance of research findings.. For some reason, this expectation of being explicit about purpose and method has not been so prevalent in traditional literature reviews and so there is a greater need to specify that a review is or is not systematic. In practice, there is a range of systematic and non systematic reviews including:

- Explicit systematic: explicit use of rigorous method that can vary as least as much as the range of methods in primary research
- Implicit systematic: rigorous method but not explicitly stated
- False systematic: described as systematic but with little evidence of explicit rigorous method
- Argument/thematic: a review that aims to explore and usually support a particular argument or theme with no pretension to use an explicit rigorous method (though thematic reviews can be systematic)
- Expert or ad hoc review: informed by the skill and experience of the reviewer but no clear method so open to hidden bias.
- Rapid evidence assessment: a rapid review that may or may not be rigorous and systematic. If it is systematic then in order to be rapid it is likely to be limited in some explicit aspect of scope.

The second term requiring clarification is 'meta analysis' which refers to the combination of results into a new product. Theoretically meta analysis can refer to all types of review but in practice the term has become associated with statistical meta analysis of quantitative data. This approach is common in reviews of controlled trials of the efficacy of treatments in health care. Statistical meta analysis is only one form of synthesis with its own particular aims and assumptions. Primary research varies considerably in aims, methods and assumptions from randomized controlled trials to ethnographies and single case studies. Similarly, synthesis can range from statistical meta analysis to various forms of narrative synthesis which may aim to synthesize

facts or conceptual understandings (as in meta ethnography) or both empirical and conceptual as in some mixed methods reviews (Harden and Thomas 2005). In this way, the rich diversity of research traditions in primary research is reflected in research reviews that can vary on such basic dimensions as (Gough 2007):

- The nature of the questions being asked
- A priori or emergent methods of review
- Numerical or narrative evidence and analysis(confusingly, some use the term narrative to refer to traditional ad hoc reviews).
- Purposive or exhaustive strategies for obtaining evidence for inclusion
- Homogeneity and heterogeneity of the evidence considered
- 'Empirical' or 'conceptual' data and analysis
- Integrative or interpretative synthesis of evidence

To date systematic reviews have only included a relatively few types of research question. Current work by the Methods for Research Synthesis Node of the ESRC National Centre for Research Methods (see, http://www.ncrm.ac.uk/nodes/mrs/) is examining the extent of variation in questions posed in primary research across the social sciences. It is then using this to create a matrix of review questions to consider possible review methods for each of these questions in order to assist the further development of synthesis methods.

Stages of a review

The variation in aims and methods of synthesis means that there is not one standard process but many approaches to reviewing. Many of these include several of the stages of reviews shown in Table 2:

Table 2 Stages of a review

(i) Systematic map of research activity

Formulate review question and develop protocol

↓

Define studies to be considered *(inclusion criteria)*

↓

Search for studies *(search strategy)*

↓

Screen studies *(check that meet inclusion criteria*)

↓

Describe studies *(systematic map of research)*

(ii) Systematic synthesis of research evidence

All the stages of a map plus:

↓

Appraise study quality and relevance

↓

Synthesise findings *(answering review question)*

This list of stages oversimplifies the diversity of approaches to reviews which do not all apply in reviews with emergent iterative methods but a brief description of each of these stages is provided here to allow some understanding of what can be involved in a review and thus the role of quality and relevance appraisal in this process:

- Review question: determining the question being asked and its scope and implicit assumptions and conceptual framework and thus informing the methods to be used in the review (sometimes known as the protocol). For example, a review asking the question 'what do we know about the effects of travel on children?' needs to specify what is meant by children, travel and the effects of travel. It also needs to be clear about the conceptual assumptions implicit in the question that will drive the methods of the review and the way that it answers the question.

- Inclusion and exclusion criteria: the definition of the evidence to be considered in addressing the question being asked. This might include, for example, the specification of the topic and focus, the types of research method, and the time and place that the research was undertaken. In a review with an emergent iterative method the inclusion criteria may not become fully clear until the later stages of the review.

- Search strategy: the methods used to identify evidence meeting the inclusion and exclusion criteria. This might include, for example, methods of searching such as electronic and hand searching and sources to search such as bibliographic databases, websites, and books and journals. Searching also varies in whether it is aiming to be exhaustive. Other strategies include sampling studies meeting the inclusion criteria, searching until saturation where no extra information is being provided by further studies, or for the search to be more iterative and explorative.

- Screening: checking that the evidence found does meet the definitional criteria for inclusion. In searching electronic bibliographic databases, the majority of papers identified may not be on the topic or other inclusion criteria for the review. For example, a search strategy on children and travel may identify studies on adult issues concerning travel with children rather than the effects of travel on children.

- Mapping: describing the evidence found and thus mapping the research activity. Such maps are an important review product in their own right in describing a research field. They can also inform a synthesis review by allowing a consideration of whether all or part of the map best answers the review question and should be synthesized by using a two stage review. For example, a map of

research on the effect of travel on children may include all types of travel and all types of effect but the in-depth review and synthesis might narrow down to examine the effect of different modes of travel to school on exercise, food intake, cognition, social mixing, and knowledge of local environments. This would exclude the effects of most long distance travel, non school travel and many other effects of travel such as safety, pollution, and self determination in travel (Gough *et. al.* 2001). The synthesis might also be limited to the types of research method thought to best address the review question.

- Data extraction: more detailed description of each piece of evidence to inform quality and relevance assessment and synthesis. The data extracted may include basic descriptive information on the research, the research results and other detailed information on the study to inform quality and relevance appraisal to judge the usefulness of the results for answering the review question.

- Quality and relevance appraisal: evaluating the extent that each piece of the evidence contributes to answering the review question. Even if a study has met the initial inclusion criteria for the review it may not meet the quality and relevance standards for the review.

- Synthesis: aggregation, integration or interpretation of all of the evidence considered to answer the review question.

- Communication, interpretation and application of review findings.

The processes of systematic reviewing are explicit methods for bringing together what we know and how we know it. This not only provides accessibility for all users of research to research findings, it also provides an opportunity for users of research to be involved in the ways in which the reviews were undertaken, including the conceptual and ideological assumptions and the questions being asked, and so provides a way of these users to become actively involved in the research agenda. This approach provides a means by which there can be greater democratic participation in the research process that is largely under the control of research funders and academics. They can also be explicitly involved in deliberative processes of involving other factors and knowledge in interpreting and applying the research findings (Gough forthcoming).

QUALITY AND RELEVANCE ASSESSMENT

Stage of review for study appraisal

In order to synthesize what we know from research, we need to ensure that the evidence is of sufficient and appropriate quality and relevance. In the stages of a review described in Table 2, quality and relevance assessment occurs between mapping and synthesis. In some approaches to synthesis the type of evidence to be included or excluded in the review might be considered to be an issue of quality and thus part of the application of inclusion and exclusion criteria in the early stages of a review.

Even if the actual process of assessment occurs at a later stage of the review process it can be considered a form of inclusion criteria. The reason for occurring later in the process may simply be because it is only after mapping or data extraction that there is sufficient information available to make the assessment. Also, when the assessment is made, it may not be an all or none decision of inclusion but one of weighting studies in terms of quality and relevance and thus the extent that their results contribute to the synthesis.

In other cases, the quality and relevance assessment can only occur later in the process because they occur at the same time as synthesis. One example is a technique called sensitivity analysis (Higgins et. al. 2006). This is a process where the effect of including or excluding lower quality studies is assessed. If the effect is minimal the studies may be included in the final results of the review.

Another example of quality assessment at the synthesis stage occurs in some of the more interpretative types of synthesis. In this case, quality and relevance assessment is an integral part of the process of synthesis, where the value of a piece of evidence is assessed according to what extra it contributes to the synthesis (for example in Realist Synthesis, Pawson 2006).

Taking all these issues together, there are at least the following ways in which the assessment of quality and relevance can occur in the process of a synthesis review:

1. Initial exclusion criteria: exclusion of types of evidence at the start of the review process: the exclusion of certain studies on the basis of their evidence type or very basic aspects of quality of the study. For example, the inclusion of only ethnographic studies or only randomized controlled trials. This narrow approach to included research designs may exclude studies with non ideal designs for addressing the review question but these excluded studies might still contain useful information.

2. Mapping stage narrowing of criteria: in a two stage review it is possible to at first include a wider group of designs and then to use the mapping stage as an opportunity to examine the whole field of research and then to maybe then narrow down to a sub-set of the studies. An alternative strategy is to include a wide group of designs all the way through to the synthesis but to use methods of quality and relevance to deal with this heterogeneity (as in ‘3’ below).

3. Detailed appraisal: detailed appraisal of the quality or relevance of the study prior to synthesis often undertaken after detailed data extraction as this provides the necessary detailed information for the assessment of studies This can be: (i) exclusion of studies not meeting the criteria and so similar to inclusion/exclusion criteria; (ii) weighted inclusion of studies assessed as non optimum on a-priori quality or relevance criteria: to allow studies to have an impact on the conclusions of the review.

4. Emergent criteria: inclusion, weighted inclusion, or exclusion of studies on basis of emergent criteria that the studies answer the review question. This is similar to a priori criteria for assessing studies (as in 1, 2 or 3) but based on emergent assessment of the contribution to answering the review question (just

as relevance of different types of data might only emerge during the process of some qualitative process studies) (Pawson 2006).

5. Sensitivity analysis: studies included or excluded on the basis of quality and relevance appraisal and the impact on the conclusions of the synthesis. Studies considered problematic may be included as long as they do not change the conclusions provided by other studies.

Weight of Evidence framework

In addition to the variation in where quality and relevance assessments fit within the stages of reviews there is the issue of whether generic or review specific quality appraisal judgements are being made.

As discussed in the first section of this paper, the concept of quality is complex and whatever the nature of the criteria applied, these can refer to more generic (or intrinsic) or more narrowly purpose and context specific judgements. A research study can therefore be assessed against more generic criteria of quality and/or against some more purpose specific criteria.

In a systematic review, a research study judged as high quality against generic criteria may not necessarily be a good study in the sense of being fit for purpose in answering the review question. The authors of the original primary study may have executed the study perfectly, but they undertook the study before the review took place and could not be expected to know the particular focus of any potential future review.

The generic form of appraisal thus considers whether a study (included in the review as meeting the inclusion criteria) is well executed, whether or not it is useful in answering the review question. Such appraisal is likely to be based on whether the study is fit for purpose in a generic way in the sense that the results of such studies performed in such ways can be trusted but it does not require any consideration of the quality or relevance of a research study for a particular research review. It is thus a 'non review specific' judgement.

Review question specific judgements consider the extent that a study is fit for purpose as a piece of evidence in addressing the question being considered by a specific review. In other words, however well executed, does the study help to answer the review question?

A first dimension of review specific quality and relevance is the type of research evidence being employed. A study may be very good of its kind but use a research design that is not powerful at answering the review question. For example, a randomized controlled trial is very appropriate for answering questions about the extent of the efficacy of interventions but unless it also has included process data it will not be not so good at answering questions of process or of the prevalence or extent of a phenomenon. Pre-post non controlled designs are not as efficient as controlled trials at addressing questions of the extent of the efficacy of interventions but are often undertaken to answer such questions due to resource constraints or other reasons (even if in the long run it may be more expensive to undertake cheaper but

inconclusive studies). On the other hand, descriptive analytic studies can be very powerful for addressing issues of process but not of extent of effect.

In some reviews there are very narrow inclusion criteria about research design so only some specific designs will be included in the review. For example, meta ethnographic reviews are only likely to include ethnographic primary research studies, whilst statistical meta analytic studies of effect may only include controlled quantitative experimental studies.

If there are not narrow inclusion criteria on research design and a wider range of designs is included for consideration, then there are issues about the relative extent that the designs of each of these studies are of sufficient fitness for purpose to be included in the synthesis in a full or weighted form.

This distinction between generic quality of execution and the appropriateness of the research design for addressing the review question avoids the confounding of these different concepts found in many available schemas and checklists for addressing research study quality. For example, a review asking a 'what works?' question about the efficacy of an intervention may only include randomized controlled trials. However, Slavin (1984, 1995) has criticized some reviews for including poorly executed randomized controlled designs whilst omitting good quality non random designs. We need a framework that allows the reviewer to make explicit decisions on these two separate dimensions of quality of execution and appropriateness of design to answer the review question. Reviewers can thus take a broader approach and include all designs and, if they wish, give less emphasis to the results of some designs over others.

A second dimension of review specific quality and relevance assessment is the topic focus or context of the evidence. Topic and context can (just like research design) be an inclusion/exclusion criterion for a review and they can also be part of the quality and relevance appraisal later in the review process. If they are part of quality and relevance appraisal later in the review, then it can be a weighted judgement allowing for a broad range of evidence to be considered that varies depending on how directly it addresses the focus of the review question. For example, a review might only include studies from the UK because studies in other countries may be undertaken in different contexts. Alternatively, the review might include studies from other countries and treat them equally or might include them and weight them lower due to the different context. Similar judgements can be made about many aspects of the studies such as the sample, the definition of what is being studied, the context and the study measures. For example, a review might want to include all the research on a topic whatever the research design being used even though those different designs may differ in their ability to answer the review question and may require different types of issues to be considered in rating their quality and relevance. As already discussed in respect of study design, a system of weighting allows for a review to employ a broader question and thus broader inclusion criteria in the knowledge that weighted judgements can be applied to the broader range of evidence identified.

Weight of evidence is a concept used in several field (including law and statistics) referring to the preponderance of evidence to inform decision making. It is a useful heuristic for considering how to make separate judgements on different generic and

review specific criteria and then to combine them to make an overall judgement of what a study contributes to answering a review question. These can create a weight of evidence framework of one generic (Weight of Evidence A), one review specific judgement of research design (Weight of Evidence B) and one review specific judgement of evidence focus (Weight of Evidence C) and an overall judgement (Weight of Evidence D) (Gough 2004):

Weight of Evidence A
This is a generic and thus non review specific judgement about the coherence and integrity of the evidence in its own terms. That may be the generally accepted criteria for evaluating the quality of this type of evidence by those who generally use and produce it.

Weight of Evidence B
This is a review specific judgement about the appropriateness of that form of evidence for answering the review question, that it the fitness for purpose of that form of evidence. For example, the relevance of certain research designs such as experimental studies for answering questions about process

Weight of Evidence C
This is a review specific judgement about the relevance of the focus of the evidence for the review question. For example, a research study may not have the type of sample, the type of evidence gathering or analysis that is central to the review question or it may not have been undertaken in an appropriate context from which results can be generalized to the answer the review question. There may also be issues of propriety of how the research was undertaken such as the ethics of the research that could impact on its inclusion and interpretation in a review (Pawson *et. al.* 2003).

These three sets of judgements can then be combined to form an overall assessment **Weight of Evidence D** of the extent that a study contributes evidence to answering a review question.

The literature contains a number of other frameworks for assessing quality of research that can be used for systematic reviews many of which can be incorporated within the Weight of Evidence Framework (see Harden, forthcoming). One example is TAPUPAS that lists seven dimensions to assess research on: Accuracy, Purposivity, Utility, Propriety, Accessibility and Specificity (Pawson *et.al.* 2003). The way in which TAPUPAS overlaps with and draws attention to issues that can be included within the Weight of Evidence framework as shown in Table 3.

Table 3. Fit between TAPUPAS dimensions and the Weight of Evidence Framework

Weight of Evidence A: Generic on quality of execution of study
Transparency - clarity of purpose
Accuracy – accurate
Accessibility – understandable
Specificity – method specific quality
Weight of Evidence B: Review specific on appropriateness of method

Purposivity- fit for purpose method

Weight of Evidence C: Review specific on focus / approach of study to review question
Utility – provides relevant answers
Propriety – legal and ethical research

Quality and relevance appraisal of reviews

The discussion so far has focused on the appraisal of individual primary research studies for inclusion in reviews. There is also the issue of the appraisal of reviews. The same Weight of Evidence framework can be used for appraising reviews as for appraising individual studies but the specific issues and criteria will vary with the aims and methods of the review. An increasing diversity of reviews is emerging and with this a range of accepted practices that will inform judgements about:

> WoE A: generic issues about quality of the execution of a review such as being explicit and transparent.
>
> WoE B: review specific issues about the particular review design employed and its relevance to the review question. For example, a statistical meta analysis might not provide much useful information about the processes of an educational intervention.
>
> WoE C: review specific issues about the focus of the review. For example, a narrowly focused review might not provide much breadth about the research knowledge relevant to answering a review question.

Such Weight of Evidence appraisals can be used for checking an individual review or for reviews of reviews.

EMPIRICAL STUDY OF QUALITY

These distinctions on quality, relevance, and weight of evidence provide a structure for making judgements but do not explain how the specific judgements should be made. In order to be systematic a review needs to specify how the different judgements of quality and relevance were made about each study and how these generic and review specific components have been combined to provide an overall judgement of what each study can or not contribute to answering the overall review question.

One strategy is a priori to define how these judgements should be made across the social sciences. Some progress could be made using this strategy but judgements of evidence quality and relevance are highly contested and progress on developing agreement might be slow.

Another, complementary strategy is to examine how people make these judgements in practice thus making explicit the often implicit ideas about quality and relevance so that these can be shared, debated and refined. This is the strategy that has been applied in the EPPI-Centre at the Institute of Education, University of London (http://eppi.ioe.ac.uk) which has supported well over twenty review groups in undertaking over fifty reviews for the Department of Education and Skills and Teacher Training and School Development Agency (see also Oakley 2003).

The majority of the reviews undertaken to date have concerned issues of effectiveness that considered experimental evidence to have most weight, although several teams did not distinguish between randomized controlled trials and quasi experimental and non controlled trials. Some review teams have stated that they give equal weight to the generic and two review specific ratings and then took an average score to rate overall weight of evidence. On the other hand, other teams have stated that they prioritised generic research quality (WoE A), and others have stated they prioritized focus of the study over other considerations (WoE C). In examining a group of reviews, Oakley (2003) reports that the review authors rated many studies to be of low overall quality.

These judgements can be considered in more detail by examining 518 primary research studies included in these reviews. For most studies (363 studies = 70%) the rating of execution of study (WoE A) and overall rating (WoE D) were the same. This suggests that the choice of method, its execution and the focus of methods were equally important to the review authors.

For nearly a third of studies (155 studies = 30%) they were different indicating that review specific issues (WoE B and C) had influenced the overall rating (WoE D). Table 4 shows that for the majority of these cases (116 studies = 73% of the 155 studies), the review specific ratings (WoE B and C) had lowered the overall rating. For the remaining studies (39 = 25% of the 155 studies) the review specific ratings (WoE B and C) had resulted in higher overall ratings (WoE D). This suggests that when review specific issues are important they are more likely to reduce than increase the overall ratings of studies. Table 4 also shows that when review specific criteria effected the overall score then this was more likely (30% compared to 14%) to be due to the effect of the relevance of the focus of the study (WoE C) rather then the choice of study design (WoE B) for both lowering (32% to 13%) and raising (26% to 15%) the overall rating.

Table 4 Weight of Evidence judgements on 155 of the 518 studies where different ratings given to WoE A and WoE

	B=C	%	B>C		B<C		Total	
A<D	23	60%	6	15%	10	26%	39	100%
A>D	64	55%	15	13%	37	32%	116	100%
Total	87	56%	21	14%	47	30%	155	100%

The next stage is to examine in detail the processes by which these review teams made and justified these assessments. This information is too detailed to be included in most summary reports but can be included in full technical reports. At the EPPI-Centre, for example, full technical reports contain specific headings to ensure that the

main methodological issues in undertaking a review are addressed including the manner in which the review teams justified their weight of evidence judgements.

This strategy of making explicit the ways in which review questions relate to appraisal of evidence enables conscious consideration of methodological decision making and fit for purpose evaluation of quality of studies in answering different review questions.

CONCLUSION

This paper opened with a brief discussion of how all different types of human activity produces different types of tacit and explicit knowledge, that is understood and used in different ways by people with very differing ideological and conceptual standpoints to develop theories and empirical statements about the world. This variation creates immense complexity for the evaluation of the quality of different types of knowledge but this diversity can be managed and understood by reference to the world views of those creating and evaluating this knowledge and their reasons for undertaking such judgements.

The paper then introduced evidence synthesis as a means of bringing together what is known in relation to any conceptual or empirical question and enabling the full range of users of research to be involved in this process. This can involve quality and relevance assessment of the research studies at various stages of a review. Despite variations in how such assessments are made there is a distinction between generic judgements of evidence quality according to generally accepted criteria (within that approach to evidence) and review specific evaluations based on the fitness for purpose of the review. The Weight of Evidence framework helps to clarify the judgements that are being used in evaluating evidence by enabling explicit decisions to be made on three dimensions of generic method, review specific method, and review specific focus and context of the study. This approach can be applied to individual studies, whole reviews, reviews of reviews, and to any quality and relevance appraisal process. Being explicit about these quality and relevance judgements then allows the empirical study of how these decisions are being made in practice so that we can assess and develop how we make these judgements. Ultimately this could provide the focus for the development of fit for purpose research methods.

In a sense, this approach is an epistemic strategy for making explicit how we identify, appraise for quality and relevance, and synthesize evidence. This should allow more open debate about how we make these decisions. This is not a strategy to mechanise or to take the value out of or in any way constrain these judgements beyond asking them to be made explicit. On the contrary, the purpose is to make the judgements more transparent so that they can be considered and debated by all. This can clarify what decisions are made on the basis of research evidence rather than other important factors such as values and resources. It can make explicit and open for debate the often implicit values and other assumptions on which the research was based and on which its quality is being appraised. It can highlight the values and perspectives behind research and encourage a greater range of people in society to engage in determining what questions are asked, the evidence used to answer those questions, and the values implicit in these processes. It can enable a democratic process both in terms of access to knowledge and also participation in its creation and use. The more

that we involve a full range of users and potential beneficiaries of research in this process, the more that we will develop a plurality of knowledge creation and use (Gough forthcoming).

Acknowledgements: I wish to thank the editors of this volume. John Furlong and Alis Oancea, and Ann Oakley and Angela Harden for their very helpful comments on earlier drafts of this paper.

REFERENCES

Furlong, J. and Oancea, A (2006) Assessing Quality in Applied and Practice-Based Research in Education: a framework for discussion. *Australian Association for Research in Education Journal*

Gough DA (2004) Systematic research synthesis. In Thomas G, Pring R (Eds): *Evidence-based Practice in Education.* Buckingham: Open University Press (pp 44-62).

Gough D (2007) Dimensions of difference in evidence reviews (Overview; I. Questions, evidence and methods; II. Breadth and depth; III. Methodological approaches; IV. Quality and relevance appraisal; V. Communication, interpretation and application). Series of six posters presented at National Centre for Research Methods meeting, Manchester, January 2007. London: EPPI-Centre.

Gough D (2007) Typology of reviews. Poster presented at NCRM Conference, Manchester.

Gough D (Forthcoming) Giving voice: evidence-informed policy and practice as a democratizing process in Reiss M. et. al. (Eds) Dimensions of Difference. London: Trentham Books

Gough D, Elbourne D. (2002) Systematic research synthesis to inform policy, practice and democratic debate, *Social Policy and Society*, 1 (3), 225-236.

Gough DA, Oliver S, Brunton G, Selai C, Schaumberg H (2001) *The effect of travel modes on children's mental health, cognitive and social development; a systematic review*. Report for DETR. EPPI Centre, Social Science Research Unit.

Harden A (Forthcoming) 'Qualitative' research, systematic reviews and evidence-informed policy and practice. Unpublished PhD Thesis, London: University of London.

Harden A, Thomas J (2005) Methodological issues in combining diverse study types in systematic reviews. *International Journal of Social Research Methods* **8**:257-271

Higgins JPT, Green S, (2006) (Eds.). Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]. Section 8.10. http://www.cochrane.org/resources/handbook/hbook.htm (accessed 22nd January 2007).

Oakley A (2003) Research evidence, knowledge management and educational practice: early lessons from a systematic approach. *London Review of Education* **1**: 21-33.

Pawson R (2006) Evidence-Based Policy: A Realist Perspective. London: Sage

Pawson R, Boaz A, Grayson L, Long A, Barnes C (2003) Types and Quality of Knowledge in Social Care. Knowledge Review 3.. London: Social Care Institute of Excellence

Slavin RE (1984) Meta-analysis in education How has it been used? Educational Researcher, 13, (8), 6-15.

Slavin RE (1995) Best evidence synthesis: an intelligent alternative to meta-analysis. Journal of Clinical Epidemiology, 48, (91), 9-18.

6564 words plus references