Systematic reviews of research in education: aims, myths and multiple methods

David Gough and James Thomas*
EPPI-Centre, University College London, UK

Systematic reviews are still a controversial topic in some quarters, with the arguments for and against their use being well-rehearsed. In an attempt to advance a more nuanced approach to thinking about systematic reviewing, this paper illustrates the wide range of theoretical perspectives, methodologies and purposes that underpin the vast range of systematic review approaches now available; and in the light of this picture, re-examines some of the perennial arguments against reviews, arguing that they are often poorly targeted, based on a misreading of what systematic reviews aim to do, or simply incommensurable with the tenets that underpin academic enquiry.

Introduction

An increasing amount of research is being published and it is difficult for professional researchers to keep up with the output, and even more difficult for those outside the academy (Bastian et al., 2010). In the past, when volumes of research were smaller and research less specialised and fragmented, the field would rely on expert opinion of what is known from previous research or on overviews of the literature: ‘literature reviews’. However, it became increasingly apparent that both experts’ opinion and the literature reviews that they produced did not necessarily reflect the diversity and breadth of the phenomena in question (Chalmers, 2003). This may not have been due to any deliberate bias on the part of the author, but owing to the fact that they were not aware of, or were unaware of the significance of, literature that was not included in their review(s). There was a clear dissonance between the rigour and transparency with which research was expected to be carried out, and the somewhat haphazard and capricious way in which overviews of research were conducted. Writers began to make the case for doing literature reviews with the same methodological rigour as would be expected for primary research: treating them as a type of research, requiring research methods like any other type of research (Davies 1999; Oakley 2003). The ‘systematic review’ developed out of this move; and, as this paper will illustrate, the approach has developed hugely over the past twenty years (Gough et al., 2012a).

This paper is in two sections. The first section concerns the characteristics of systematic reviews, aiming to highlight the breadth of perspectives and approaches that the term now embraces, and how systematic reviews can be used to inform
decisions. In the second section of this paper, we move to address, more briefly, some of the recent history and debates around systematic reviews over the past 20 years, including concerns about governmental control of the research agenda, prioritising randomised controlled trials in education and privileging positivist approaches to educational research.

The breadth of approaches used in systematic reviews

What is a systematic review?

If one searches online for definitions, a variety of offerings come to hand with most coming from clinical research (Akobeng, 2005). They all emphasise different aspects of systematic review methods depending on the situation for which they have been prepared. While most definitions foreground the importance of the research question(s) to be addressed, it is clear that there is no one definition for ‘systematic review’ and there is no one template that all reviews follow, though there is some common terminology (Cooper et al., 2009; Gough et al., 2012a).

Many descriptions of systematic reviews begin by outlining the ‘stages’ a review goes through, from identifying an answerable question, to developing inclusion criteria and search strategies, to methods of quality assessment and synthesis (Gough et al., 2012a). These descriptions tell us that their authors see a systematic review as a piece of research, and like any primary research, are conducted using an established, though emerging, research method. They also describe some of the common components of systematic reviews: ‘search strategies’, ‘inclusion criteria’, ‘quality assessment’ and ‘synthesis’, to name but a few of the more common terms. While giving us some insights into how some reviews are done though, these descriptions do not really tell us what a systematic review is or why these stages are deemed necessary.

Systematic reviews originally developed as a new way of doing literature reviews—one that saw the utility of reviews of the literature—but which addressed criticisms of the way they had traditionally been conducted (Oakley, 2003). They have two main principles. They aim to avoid drawing wrong or misleading conclusions that might arise from:

1. Possible deficiencies in the studies they include.
2. Possible deficiencies in the way that the review itself is carried out.

One of the criticisms levelled at traditional literature reviews was that readers could not tell where the included studies came from and how they came to be included. To address this, systematic reviews aim to be clear about what research is, and is not, relevant—and why. Clarity is therefore sought with regards to where and how searches for research were conducted; aiming to be transparent in their reporting of these—and other issues—so that the reader can judge for themselves how well conducted the review was (Chalmers, 2003). (Complete transparency is, of course, impossible, but systematic reviewers accept that they need to be accountable for the way that their reviews are carried out.)

In order to avoid drawing conclusions based on research that is itself unreliable, the studies that are included are usually assessed for their quality and rigour in some way.
Reviews adopt different strategies with regard to this, depending on their priorities for quality. Some will include or exclude studies based on their research design, whereas others will include all studies and determine their quality based on their relevance and ability to inform the review’s findings. Being critical about the literature included, however this is carried out in practice, is expected practice in systematic reviews (Gough et al., 2012a).

Beyond these principles, there is considerable diversity—as in primary research—in the detail of how systematic reviews are conducted in practice. They are not a homogenous entity, and serious discussions of their practice and potential need to allow for this diversity.

The above principles—that reviews should be clear about what they did, and that they should only base their findings on reliable research—might appear to be uncontroversial and markers of any good literature review, and this is really the main point of this paper. Systematic reviews take markers of how good research is carried out and apply these principles to the process of conducting reviews. Good literature reviews follow the above principles, though of course many did (and do) not. While there are legitimate and complex discussions to be had as to the fitness of the purposes to which systematic reviews can be put, these should not cloud our assessment of the fitness for purpose of the kaleidoscope of methods now emerging.

Mapping and synthesis in reviews

In undertaking a review, researchers capture information about the findings of each included study in the form of ‘categories’ or ‘codes’ in order to be able to synthesise the data. They also create categorisations that encapsulate dimensions of study quality in order to be able to appraise the strength and relevance of those findings; and a variety of other descriptive information about a study. Initially this information was largely seen as bibliographic and descriptive, but in the late 1990s, colleagues at the EPPI-Centre saw the benefit of describing the characteristics of studies in a review as a product in its own right; as a description, or ‘map’ of a research field (Peersman, 1996). What was described depended upon the interest of the reviewers, but could be anything from methodological descriptors to the theories behind the phenomenon being studied. The map could show what had been studied and how it had been studied and so could identify methodological and topic gaps in research effort.

Maps also inform how studies are combined later in the review (the ‘synthesis’). A map might show that the included studies are too heterogeneous for synthesis and so suggest that only one or more sub-parts of the map lead to synthesis. This then allows for broader research questions to be asked at the start of the review in the knowledge that the review’s scope can then be narrowed at the synthesis stage. One benefit of this approach is that a greater range of research methods can be included at the mapping stage, ensuring that any exclusion criteria based on study design can be selected in the knowledge of the likely range of research available. Any later exclusion of studies with certain methods then becomes an overt process with all potentially relevant studies described, rather than being hidden if excluded totally from a review.
Methodological approaches in reviews

In primary research there are very many methods of enquiry depending on the theoretical and ideological perspective and the specific research questions being asked. All of this variation in primary research questions and methods is replicated at the review level. Systematic reviews tend to use methods similar in approach to those used to answer the same questions at the primary level and to include those types of primary studies in their review. However, the selection of method and purpose is a complicated process and while approaches tend to cluster together, thinking in terms of the traditional qualitative–quantitative divide only takes us so far. Instead, we have found it helpful to consider that reviews differ in terms of continua; that is, they can prioritise one approach over another, but they can rarely be completely pigeonholed at one or other end of any given spectrum. Figure 1 aims to encapsulate this way of thinking, depicting two broad clusters of approach (the vertical arrows) along with horizontal axes that highlight significant ways in which approaches vary.

The traditional image of a systematic review, a review on the effects of interventions evaluated in experimentally controlled trials would be located within the right-hand column. These reviews usually adopt a realist philosophical stance, in that they assume that there is an observable world which reviews aim to understand a little more about through the testing of hypotheses, the identification of a set of relatively homogenous studies and the aggregation of their findings (Gough et al., 2012b). Aggregation is a key concept in synthesis, since it encapsulates the increase in certainty and precision that is thought to arise from multiple observations of the same phenomenon; i.e. we find many studies that evaluate the same intervention, pool their results and have greater confidence in the magnitude and precision of the pooled ‘effect size’ than we would have if we looked at a single study alone.

As an example of a review type that would be located within the left-hand column, a meta-ethnography includes qualitative research and uses methods of review similar to those found in primary qualitative research (Britten et al., 2002; Noblit & Hare, 1988). In these reviews, philosophical stances can move towards the idealist end of
the spectrum, where the aim is to understand people’s perspectives and experiences in order to generate theory through the configuration of study findings. Configuration is another key concept in synthesis, encapsulating the action of placing study findings alongside one another in order to build up a picture of the whole, and how they relate to one another, i.e. configuration can help us understand diversity, explaining why a given situation pertains in one context but not another (Sandelowski et al., 2012).

Reviews on the effects of interventions are concerned about bias and aim to be exhaustive: they usually try to include as much as possible of the extant research relevant to the question they are addressing in order to avoid their conclusions being based on an incomplete picture of the research available. They also tend to have an a priori statement of their methods in order to avoid bias creeping in when the results of studies are known, and emphasise the importance of assessing the methodological quality of the studies that they contain (Cooper et al., 2009). Primary research on the effects of interventions also tries to avoid such bias by studying total populations or at least representative samples of that population and the pre-registration of trial protocols is now an accepted means of reducing bias in trials and increasing methodological rigour. In contrast, meta-ethnography is less concerned with bias and the representativeness of primary samples or the exhaustive inclusion of studies in a review and more concerned with the range of examples included. In this case, searches are often iterative and aim to identify a broad range of perspectives, rather than requiring that every single relevant study be located. The ‘quality’ of the included studies relates both to the contribution that they are able to make to the emerging theory and how well they were carried out (Britten et al., 2002; Noblit & Hare, 1988).

The aims of reviews can also differ. Weiss made the distinction between the instrumental and enlightenment use of research (Weiss, 1979). The aggregative review of intervention effects can tell us whether or not a given intervention ‘works’, providing us with an estimate of the magnitude and precision of its effect. Such a review aims to provide decision-makers with knowledge that can be used instrumentally, e.g. ‘should we invest in intervention A or B?’. In contrast, the configurative meta-ethnography creates new theory, providing us with new ways of understanding people’s beliefs and experiences. These reviews enlighten the perspective of their readership, aiming to influence decision-making through changing their perceptions and opinions, rather than giving them a straightforward answer to a precise question.

On a first read, the configurative–aggregative distinction appears to map quite closely onto the traditional qualitative–quantitative divide, but because this is a logic of synthesis, such a mapping does not hold in practice. The following examples will explore this logic.

**Effectiveness questions (‘what works?’)** We have already mentioned that effectiveness questions are a common example of a predominantly aggregative quantitative form of research using statistical analysis to test a hypothesis (that the intervention under study does work). A common distinction between studies (or parts of studies) taking an aggregative or configurative approach is the extent that theories and concepts are decided prior to the study. In meta-analysis, where study findings are combined statistically, the main objective is to test a pre-determined hypothesis using a pre-determined method. Iteration is avoided for fear of bias from the analysis being
data-led. However, post hoc exploration of the relationship between variables in (meta-) regression analysis is more iterative and more configurative. Here, findings are aggregated within given covariate (categories) and configured between them, and the question then becomes ‘how does the phenomenon vary, depending on another phenomenon’, rather than ‘what works?’

Another paper in this issue looks in more detail at methodological developments in reviews that utilise statistical analyses to synthesise study findings (O’Mara-Eves and Thomas, 2016). Many of these developments are concerned with understanding more than simply the effect of individual variables. They are concerned with how variables relate and in the causal mechanisms explaining these relationships. They are concerned with explanation and theory and so tend to have more complex hypotheses and more post hoc configurative iterative exploration. These reviews are therefore not quite so far to the right of the dimensions in Figure 1 as a straightforward meta-analysis of intervention effects, as they contain both aggregative and configurative aspects.

One interesting methodological development is ‘living reviews’ (Elliott et al., 2014b). These are reviews that are ongoing and are updated whenever a new study is reported and then added to the statistical analysis. This can change the way in which primary research is planned and undertaken. In deciding upon sample size in a primary randomised controlled trial (RCT), a power calculation is undertaken to find the necessary sample size required to be able to show the hypothesised effect with the outcome measures being considered. With living reviews, however, the power calculation can be made on what sample size would be necessary in a new study to be able to change the conclusions of the ongoing living systematic review. Living reviews change the way that we distinguish primary research from reviews; they are in a fluid dynamic with each other and both part of an ongoing strategic approach to developing questions, developing theories and developing data produced from those theories.

A related issue is the relationship between data from individual research participants in studies and a systematic review of studies. In the past, reviews were based on the synthesis of summarised data from each primary study but now techniques are being used to synthesise individual level primary data (Riley et al., 2010). Data on individuals may also be used to contextualise and interpret and apply the findings of reviews in different contexts or for different individuals. Primary and systematic review findings are generalisations and data about the range of individuals in society (from, for example, administrative data) or about one specific individual about whom a decision is about to be made can assist in fine tuning that decision.

These developments reflect the way that the distinction between RCTs and longitudinal large scale routine data is narrowing. In the past, a RCT often required expensive data collection of participant outcomes. With the development of the national pupil database in the UK, outcome data on student attainment can be accessed for little cost. As more data becomes available on more issues, the relationship between individual data points, primary research and reviews of that research are beginning to merge (Elliott et al., 2014a).

Another issue is the assessment of cost effectiveness as well as simple effectiveness. If two interventions, for example, ways of teaching foreign languages, are equally effective, but one is much cheaper than the other, then it is more cost effective. In practice, cost effectiveness reviews are difficult to undertake in education as so little
information is collected in a rigorous or consistent way in primary studies on cost (Shemilt et al., 2013). If such information is available, it does not only allow decision-makers to choose the most cost-effective solution to a particular service need, it can also allow for resource allocation decision between very different services. In health services in England, for example, the National Institute for Health and Care Excellence has created the concept of a quality adjusted life year (QALY) to provide a metric for comparing how much a health intervention costs in relation to producing one good quality of life year (National Institute for Health and Care Excellence, 2010). It would be possible to develop similar metrics in education—for the cost of one good quality year of education or learning. This would of course require (as is the case in health) much debate about the values involved in making such assessment of quality of education or learning (Gough & Kenny, 2014; Kelly et al., 2015).

**Configurative reviews to develop theory** In contrast to the aggregative focus of ‘what works?’ reviews of trials, other reviews ‘configure’ findings in order to develop new concepts and theory. Meta-ethnography is a common example of a predominantly configuring qualitative form of systematic review that uses meaning and interpretation to construct new concepts and theories. This approach is much more iterative and exploratory than the a priori hypothesis testing of ‘what works?’ reviews. Although these qualitative studies are predominantly configuring concepts, they may make some summative comments and thus involve a degree of aggregative analysis (Britten et al., 2002; Noblit & Hare, 1988). Another paper in this issue looks in more detail at methodological developments in qualitative reviews (Nye et al., 2016).

One important way in which these predominantly configuring qualitative reviews vary is the degree that the concepts are developed iteratively. In thematic synthesis (Thomas & Harden, 2008) and meta-ethnography (Noblit & Hare, 1988), there is only a very general framework provided by the review question. Concepts and theories are developed (or constructed) iteratively during the process of undertaking the review. In framework synthesis (Oliver et al., 2008), on the other hand, there is a pre-existing detailed conceptual framework that is modified during the process of the review. An analogy would be the extent that questions in a questionnaire survey provide a framework of pre-set answers to select from or that the answers are left open-ended for the respondent to develop. In sum, the degree of iteration and development of concepts varies between reviews and/or between the stages of a review (as in the a priori and exploratory stages of some ‘what works?’ reviews).

Another and highly related issue is the extent of iteration of method (rather than concepts) during a review. In a standard ‘what works?’ review, the method is specified a priori and any variation is considered a potential source of bias except for post hoc exploration of the data (which should come with warnings about the exploratory nature of the exercise). In many reviews that develop theory the method is also often pre-specified but in some, such as realist reviews, there is considerable flexibility in how parts of a review may progress (Pawson, 2002). [Of course, it is possible to pre-specify the fact that an iterative approach will be adopted (Barnett-Page & Thomas, 2009).]

**Mixed methods, multi-component reviews, complex reviews and complexity** In primary research, many broad or complex questions can be investigated by more than one method, so many advocate mixed methods primary research (Creswell & Plano Clark, 2011). Again the logic of primary research can be applied to systematic
reviews. Mixed methods reviews often consist of two or more sub-reviews addressing different aspects of the overall review question and so can be thought of as multi-component reviews (Harden & Thomas, 2010; Sandelowski et al., 2012; Voils et al., 2008).

One reason for using multi-component reviews is simply that the review question is broad and the literature on this topic is best analysed as separate components (or sub-reviews) before being combined in an overall synthesis. A second reason is that the review question raises both empirical and conceptual issues and these require different review methods and probably also different primary studies to be included. A review might thus have a ‘what works?’ component with a predominantly aggregative analysis and a conceptual synthesis with a predominantly configuring analysis. The two arms or components are then together to provide a more holistic understanding of the research evidence available (Thomas et al., 2004). It may be that the configuring synthesis is used to further understand an aggregative synthesis, or vice versa; or they are both used to examine the other synthesis (Heyvaert et al., 2011; Pluye & Hong, 2014).

Many applied social research questions, including educational questions, are highly complicated in terms of both theory and data. Many of them are also complex in that the variables being considered may interact together in more than simple additive ways (Noyes et al., 2013; Thomas et al., 2014). This may govern the extent that a single systematic review, even with mixed methods and multiple components, may be limited unless there has been a prior period of conceptual and empirical work bringing the state of research to a point where such a review can be helpful.

Maybe a better way to understand the role of any review is as one part of a longer term strategy to further our understanding on an issue. At different points in time there will be methodological and pragmatic challenges that lead researchers to explore different avenues or to need particular types of primary or review study to progress understanding. So systematic reviews have an important role in clarifying what is known and not known about an issue considered from a particular perspective at one particular point in history, but they are also important stages within the research process itself. Systematic maps can inform us about what has been studied and how, and thus identify gaps in that research investment. Systematic synthesis can clarify what is known but also feeds back into what more needs to be known.

It is also interesting to consider realist and meta-narrative reviews as a particular type of mixed methods review. Realist reviews, like realist evaluation in primary research, are concerned with what interventions work, but also for when, where and how (Pawson, 2002). They are concerned with explanation of how things work as well as whether they do work. Other primary research and systematic reviews also have these concerns. What seems different about realist synthesis is the use of iteration in method during the process of the review. The first stage of a realist review can be seen as the unpacking of theory underlying a review question and this is mainly a configuring form of analysis. The second stage is the empirical testing of the different parts of the theory or theories that have been identified. In standard approaches this empirical testing stage would be using detailed a priori methods, whereas in realist synthesis it is more of an investigation following whichever paths open up to assess whether the necessary conditions, causative agents and outcomes are present to indi-
cate empirical support for the theories implicit in the review question being asked (Gough, 2013a).

Greenhalgh and colleagues have described a ‘meta-narrative’ approach to systematic review, based on Kuhn’s *The Structure of Scientific Revolutions* (Greenhalgh et al., 2005). This method is grounded on the principle that groups of researchers view the world through their own ‘lens’ and produce work that is incommensurable: that ‘empirical discovery made using one set of concepts, theories, methods and instruments cannot be satisfactorily explained through a different paradigmatic lens’. Greenhalgh and colleagues identify relevant literature, group studies according to their disciplinary roots, and then construct a series of ‘meta-narratives’ through which they are able to chart emerging schools of thought, where differences in results are explored through differences in disciplinary perspective.

In sum, reviews are a level of analysis rather than a fixed method. There are some specific methodological issues concerning reviews, but in general, the methods reflect the question being asked and the methods (and the studies using these methods) used to answer these questions. The questions and the reviews used to answer them may have sub-questions and sub-reviews. These feed back into further primary research, making reviews less of a one-off activity and more part of an ongoing developmental strategy of researching different issues using primary research and reviews of that and related research. This flexible and dynamic, knowledge-cumulating perspective of reviews, fits with the earlier discussion of the development of complex reviews and multiple levels of data and theory for reviews of effectiveness. Research is less of an event than a process with many levels of data analysis. Systematic reviews are just one level and one part of such a data and analysis system.

Information science In addition to methods development, there are also major developments in information technology that impact on reviews. Improvements in technology and the ever-increasing ‘data deluge’ are prompting increased use of information technologies to support reviewing. With more and more research being published, it is becoming increasingly difficult to find and utilise knowledge; the task of separating the relevant from the irrelevant is simply requiring too much time.

To help with this problem, automation is beginning to develop to assist the review process. New, federated search engines are being developed, which search multiple sources at the same time, eliminating duplicates automatically (Tsafnat et al., 2013). Technologies, including text mining, are being used to help develop search strategies, by suggesting topics and terms on which to search—terms that reviewers may not have thought of using. Searching is also being aided by technology through the increased use (and automation) of ‘citation chasing’, where papers that cite, or are cited by, a relevant study are checked in case they too are relevant.

In the past, much review work was undertaken using bibliographic management software, but there is increasing use of specialised review software to support and in some cases automate the review process. Text mining, for example, can assist in screening studies for a review as well as with the task of searching (Thomas et al., 2011; Wallace et al., 2012). The process of identifying studies that will help answer the review question (meet the review ‘inclusion criteria’) is not efficient, and searching bibliographic databases leads to many irrelevant studies being found which then need to be screened (checked) manually one by one to find the few relevant studies. A
typical text mining or machine learning process might involve humans undertaking some screening, the results of which are used to train the computer software to learn the difference between included and excluded studies and thus being able to indicate which of the remaining studies are likely to be relevant or not. Such automated supports may result in some errors in screening, but these errors may be fewer and less important than the current errors produced by humans manually screening thousands of mostly irrelevant citations (O’Mara-Eves et al., 2015).

Likewise, technologies are being developed that are able to classify studies according to their key characteristics. Data-driven ‘clustering’ technologies are available, which analyse the text of documents and group similar papers together. These technologies can be used to ‘map’ (see earlier section for the use and definition of maps) research quickly with minimal human input (Stansfield et al., 2013). It is also possible to classify research according to pre-specified criteria—such as controlled vocabularies and thesauri; and it is now possible to automate the assessment of the ‘Risk of Bias’ of a RCT according to specified tools (Marshall et al., 2014).

Though not yet pursued to the same extent in education, the automated curation of health knowledge is now reaching a point whereby many of the laborious, manual tasks that have traditionally made up the bulk of reviewing labour can increasingly be undertaken by machines. Systems are being established that automate the identification of research, classify it according to domain, and make it available for reviewers as and when it is needed. Automation is overlapping with citizen involvement in science in this area, and ‘crowds’ of people are getting involved in the curation of knowledge, where their effort is focused and made more efficient through the involvement of machine automation (Thomas et al., 2015).

Formalising review methods and processes As the methods and application of systematic reviews develops, there is an increasing plurality of types of reviews. This plurality reflects the energy and growth in this area and, as previously mentioned, creates the same diversity in reviews as exists in primary research. There have been three main consequences of this diversity.

First, there has been a proliferation of methods and with new names to describe them. As others apply and develop these methods there is not always consistency in the detail of the method. In a sense, what it says ‘on the tin’ does not always reflect what is undertaken inside.

Second, in order to maintain rigour and quality, a range of standards have been produced on either how to undertake reviews or reporting guidelines that have a similar effect in specifying what a review needs to contain and be reported on. Examples include Methodological Expectations of Cochrane Intervention Reviews (MECIR) standards (Chandler et al., 2013) and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement as to what a review report should contain (Moher et al., 2009).

A third consequence is increased diversity in the planning and reporting of the review question and the components of the methods used to answer that question. Reviews vary in terms of their type, breadth and depth of question, review components, and aggregative and configurative modes of synthesis. In the absence of clear holistic review types across systematic reviews, describing the dimensions of difference clarifies the aims and methods being applied and enables any quality and rele-
Vance assessment of the reviews (Gough et al., 2012b). It encourages a fit for purpose approach to developing the methods used to answer particular questions in a review.

**Using systematic reviews to inform decisions**

This paper, like many before, has argued that we need to have formal methods for bringing together findings from different research studies (Davies, 1999). The argument is that this will assist in knowing what we know and do not know and what further we may want to know through future research.

This argument assumes several things. First, that there is relevant evidence to be synthesised and made available to decision-makers. Different individuals and groups in society have different priorities and assumptions that drive the questions that they might ask of research, and research may not always fit those needs. The existence of many studies on a particular topic does not mean that there is necessarily relevant (or good enough quality) research to answer particular review questions. Second, the research needs to use fit for purpose methods appropriately so that the research findings are trustworthy. Third, the research needs to be synthesised using fit for purpose questions and methods. In some areas many systematic reviews have already been conducted, so fit for purpose reviews of reviews may also be needed.

A fourth assumption, or necessary component, is that the findings of the synthesises are available and communicated to those who make decisions. In the UK in education there have been several initiatives to enable this. The government from 2000 to 2010 supported systematic reviews and the summary of research findings on government or government-related websites. More recently, the Education Endowment Foundation has provided a toolkit providing summaries of what is known about the effectiveness of a range of different education-related interventions (Education Endowment Foundation, 2015). They have also commissioned a number of research studies to assess the most effective way of communicating messages from research findings to school teachers.

This leads to a fifth component of how research is interpreted to inform decision-making. With web access to research summaries and toolkits, the emphasis is usually on the reader interpreting the relevance for their own situation. However, in other fields, such as health, there is more emphasis on providing guidance: an interpretation of the relevance of the findings to what the reader might decide based on that evidence. The extent of the guidance ranges from relatively context free recommendations [as in the GRADE process (Guyatt et al., 2008)] to formal guidance in National Institute for Health and Care Excellence guidelines. The guidance may also be so specific that it states how such decisions should be implemented in practice, which is the sixth component of the process. This assistance with communication, interpretation and implementation of evidence may be through written materials but it may also be more interactive using humans to be knowledge brokers between producers and users of research.

In sum, there are these multiple stages of primary research production, synthesis, communication and interpretation, and implementation of research findings. All of the factors discussed so far are very much driven by the production (push) of research rather than demand (pull) from the potential users of research. Demand approaches...
start with the potential use of research to inform decisions—there is an issue being considered and research is examined to see if it can help inform such decisions. Such approaches are not easily separated from the production process, as without the primary research there are no findings that may be informative to decision-makers. However, a demand approach may change the nature of the evidence to use system. The demand can provide decision-maker ‘user’ perspectives in the commissioning of research, in reviews of that research and in the way that research findings are made available (Oliver et al., 2004).

Such approaches can also change the expectations of others in the use of research. One example is provided by the UK Parliamentary Select Committee asking the Department for Education for information about the research that informed their future government policy plan (Commons Select Committee, 2015). The simple asking of that question may prompt the Department for Education to be more explicit about how it seeks and considers evidence in policy-making. This reveals a seventh part of the process of evidence use, a feedback loop from demand for evidence that not only seeks reviews of research, but also feeds back into the priorities in deciding upon primary research. Systematic reviews provide a means by which the demand from research can feed back into that prioritisation of research investment.

This way of looking at the production and use of research as a dynamic system also relates to the earlier discussion on the changing types of reviews and changing relationships between primary research and reviews. We have looked at the way in which reviews are:

- a level of analysis rather than a fixed method;
- less of an event than a process with multiple levels of analysis;
- linking in dynamically with ongoing administrative data and large data sets;
- varying in providing summarised versus individual data;
- varying in detail on the context specificity of findings; and
- on a continuum between research findings and guidance derived from those findings.

In light of this, the demand pull for research to inform decision-making is less of a mechanistic adoption of specific results but more of a dynamic engagement between decision-making and varying sources of research-based information.

Research on research use In addition to the concern that research should have an impact and that it should be used to inform decision-making, there has been a recognition that we need to study these processes; in other words, that to know how best to use research we need to study the mechanisms and impacts of different evidence use strategies and to undertake research on research use (Nutley et al., 2007; Weiss, 1979). This is a new field of study across many disciplines of research in terms of the research that is meant to have an impact and in terms of the disciplinary approaches used to study the impact. Currently, we know a lot about the barriers to research use, quite a lot about necessary conditions such as relevant research that is communicated well, but rather less about the sufficient conditions to make evidence use happen in practice (Gough et al., 2011).

Another strategy has been to put the emphasis on implementing the recommendations of evidence synthesis rather than attempting to use evidence to persuade others.
of the importance of evidence in making their own decisions. In public health, for example, implementation science includes behaviour change strategies that try to influence people’s health behaviour. The desired change in behaviour may be evidence-informed, but the people who are having their behaviour changed may be unaware of this.

The behaviour change approaches are considering the implementation of agreed practices. This is complicated enough; the ways that research informs decision-making makes this even more complex and difficult to study. It includes the dynamic systems of research production and use of primary research and reviews plus the myriad political and rational cognitive factors involved in decision-making.

Recent history and debates around systematic reviews

Why are systematic reviews sometimes seen as controversial?

The idea of doing a literature review well and ensuring that it is a genuine reflection of the state of knowledge about a given subject, appears on first sight to be fairly common-sense, and uncontroversial; why would you want to do an unsystematic review? However, even if the idea of a systematic review is simply ‘doing a good literature review’, like all knowledge-seeking—and creating—endeavours, the way in which you go about it, and the purposes for which you do it, can be seen as political actions, which by their nature, can be seen as promoting one epistemology over another.

As mentioned in the introduction, systematic reviews have been criticised in some quarters and are still a controversial topic to some in the education field. In the light of the rich diversity of review activity that we outlined above we think that many of these criticisms are either out of date or exaggerated. Even so, it is perhaps worth revisiting the recent history of systematic reviews in education briefly and, in particular, the anxieties about governmental control of the research agenda, randomised controlled trials and fears of positivism. The following sections help to illustrate that the main criticisms of systematic reviews in education generally concern wider issues about the nature and purpose of educational research.

Governmental control of the research agenda

Historically there have been many calls for science to balance the investment of energy into primary research with the synthesis of what is already known. The James Lind Library provides an illustration of such arguments from Hippocrates arguing for research to build on the work of others in the fifth century BC, to James Lind studying scurvy in the eighteenth century, and to the rapid increase in methods of review in the second half of the twentieth century (James Lind Library, 2015).

In addition to the many individual calls for more formal methods of synthesis, the formation of the Cochrane Collaboration in 1993 provided a structured forum for the execution of, and development of methods for, ‘what works?’ reviews in health care. It developed an international network of contributors focusing on specific topic areas and methods specialists to monitor and develop review methodologies. The Cochrane Collaboration, and the broader vision of using evidence to inform policy and practice
decisions, received much support from the UK government that provided funds to the UK Cochrane Centre and created an overarching system for the production, synthesis and use of research in what is now the National Institute for Health Research.

When the new Labour government came to power in 1997, it emphasised the importance of evidence to inform and monitor policy decisions, and built this into their plans for ‘Modernizing Government’ (UK Government, 1999) and into the growth of evidence use policies in education (Gough, 2013b). The Department for Education and Employment (now the Department for Education) commissioned the EPPI-Centre to recruit and support groups to undertake systematic reviews on various areas of education. The aim was to develop capacity and review products to inform policy-making (Oakley et al., 2005). The government also set up an independent body, called the National Education Research Forum (NERF) (set up in 1999 but the main practical work started in 2002), with the aim to make evidence in education more available and to encourage a more strategic link between development and research involving teachers, policy-makers and researchers. NERF worked with funders and journal editors, and the forum also discussed research needs in specific topic areas. NERF proposed to government the creation of a National Evidence Centre for Education but this was not supported and NERF came to an end in 2006.

The UK government’s initiatives in NERF and systematic review groups occurred during a period of criticism of educational research and the lack of use of research evidence in educational practice (e.g. Hargreaves, 1996). These critiques and developments were not well received by some academics in education (Ball, 2001; Hammersley, 2001). There were at least three concerns. The first was that an interest in evidence-informed policy was an aspect of a new managerialism where there was a seeming technocratic process but this hid particular ideological and theoretical perspectives, so it was in effect a way of managing the educational research agenda. The defence to this was that both the systematic review groups and NERF were independent of government with high levels of stakeholder representation. Systematic review topics were, for example, not determined by government, though in practice this made them less useful for informing policy.

A second and related concern was that the role of academia should be to critically appraise the action of governments. There was anxiety of danger for academics of their work being overly shaped by government’s applied needs (Whitty, 2006). We, however, believed that the transparent identification of research questions and their underlying perspectives and the sharing of synthesised research findings would be more open to democratic participation and debate than the traditional academic selection of research agendas and reporting of the evidence from single studies (Gough & Elbourne, 2002). Moreover, systematic reviews themselves can be important ways of critiquing the evidence base for policy decisions.

Concerns about privileging randomised controlled trials and positivist approaches

A third concern was that systematic reviews privileged RCTs over other forms of research. At that time, RCTs were very rare in UK educational research and there was an anxiety that they would become popular. As the questions being asked by
reviews were predominantly of the effectiveness of different educational strategies, most of the reviews were seeking experimental evidence. However, the proposed NERF research agenda of development and research was broader than RCTs and, as has already explained, there is nothing in the logic of systematic reviews that limits them to such questions and methods.

At the same time as the development of an interest in the synthesis of experimental studies in the UK, there was a growing interest in experimental research in the USA with the policy of ‘No Child Left Behind’ which focused Federal funding of State initiatives on programmes that had evidence of effectiveness. A “what works?” clearing-house for education was also set up which had close links to the Campbell Collaboration, a sister organisation to the Cochrane Collaboration for education, social care and crime, that was originally based in the USA.

In the UK, the initial enthusiasm for developing a programme for evidence synthesis in education waned but was revived with the new Coalition government in 2010 which provided a large founding grant to the Education Endowment Foundation (EEF) with a remit to evaluate what interventions work in breaking the link between family income and educational achievement. EEF thus created the first large programme of systematic reviews in the country. It also created a toolkit to provide user friendly summaries of research syntheses on a wide range of educational topics to enable discussions by policy-makers and practitioners about the decisions they make about educational provision in schools (Education Endowment Foundation, 2015). The government has also launched a number of different ‘what works’ centres across different areas ranging from wellbeing to crime with the EEF included as a pre-existing what works centre.

In the case of the systematic review, methods and practice have developed particularly rapidly over the past 25 years in the field of health research, where the priority has been to identify which treatment option is most appropriate in a given situation. Systematic reviews in that field have therefore focused on synthesising the findings of controlled experiments to assess the impact of interventions, as the RCT is considered to be the most reliable way of determining whether an intervention did indeed have the desired effect. RCTs have been frequently challenged on epistemological grounds, especially within education, with some claiming that they signal a return to an empty empiricism and naïve positivism that has long-since been discredited. Whether or not RCTs can legitimately be described as positivist, the fact that early developments in systematic review methodology focused on RCTs has led, in some quarters, to the terms becoming almost synonymous with one another, and the charges of positivism being levelled at systematic reviews as well. This is unfortunate, as the logic of using systematic research methods applies to all research questions and all methods of primary research, and the negative attention paid to one, to the exclusion of the rest, has acted as a brake on epistemological and methodological progress.

The other bone of contention to highlight at this juncture is the fact that many systematic reviews are conducted to inform decisions: they are one of the main methods promoted by those advocating for evidence-based policy and practice. Again, this touches nerves in certain quarters for different reasons, the role of the academic—and academic knowledge—in society being foremost, but the aforementioned concerns
about governmental control of the research agenda also present. There are interesting discussions to be had in relation to these concerns too, but like the above conflation of concerns about RCTs with concerns about systematic reviews, these apprehensions are broader than the question of how one should go about doing a systematic review and therefore lie outside the scope of this paper.

Conclusion

This paper has outlined some of the major epistemological approaches to systematic reviews, highlighting the breadth of approach and also the emerging foci on new types of data, technologies, and the way that research into the use of research is developing into a major field of enquiry in its own right. More detail on some of the main approaches can be found in the other papers in this issue. Although systematic reviews grew rapidly from particular questions, perspectives and methods, this is now history; we now have a broad and growing range of approaches at our disposal. We have therefore also used this high level overview to show that many of the arguments advanced against systematic reviews as a legitimate and useful way of understanding knowledge in an area are poorly targeted, being based either on an out of date understanding about the breadth and potential of the approach, or are really concerned with wider issues about the nature and purpose of educational research. There are many papers and books that go beyond the scope of this paper and give the positive case for why systematic reviews are both practicable and important.

The need to understand a field of enquiry is a basic issue in research and systematic review methodology has been constantly advancing in order to encompass the breadth of approaches, epistemologies and questions that primary research itself covers: the idea of systematic review is not owned by one research method or paradigm. Similarly, research questions are not owned by one perspective or world view such as government, ‘lay’ people—or academics. Reviews which examine research and knowledge from multiple perspectives facilitate the inclusion of different voices, providing a means for non-specialists to engage with research and thus to be engaged in research debates (and to challenge those who provide non-transparent accounts of what they report the research literature to say).

We have also moved away from the idea of a review being a single entity and type at one point in time, to being something that can be understood as being much more dynamic; involving the critical engagement with data and knowledge claims, in the light of how these claims originated—and what their implications might be in different contexts—in order to inform decisions which affect people’s lives.

References


Thomas, J., Harden, A., Oakley, A., Oliver, S., Sutcliffe, K., Rees, R., Brunton, G. & Kavanagh, J. (2004) Integrating qualitative research with trials in systematic reviews: An example from public health, *British Medical Journal*, 328, 1010–1012.


