Reflections on the Methodological Approach of Systematic Reviews

Martyn Hammersley

1 Introduction

The concept of systematic reviewing of research literatures became influential in the second half of the 20th century, in the context of the longstanding, and challenging, issue of how to 'translate' research findings into reliable guidance for practical decision-making—to determine which policies, programs, and strategies should (and should not) be adopted (Hammersley 2014; Nisbet and Broadfoot 1980). The idea that research can make a significant contribution in assessing the effectiveness of policies and practices was hardly new, but it was greatly bolstered around this time by the emergence of the evidence-based medicine movement. This identified a problem with the effectiveness of many medical treatments: it was argued that research showed that some commonly used ones were ineffective, or even damaging, and that the value of a great many had never been scientifically tested; despite the fact that such testing, in the rigorous form of Randomised Controlled Trials (RCTs), was feasible. Subsequently, the idea that practice must be based on research evidence about effectiveness spread from medicine to other areas, including education.

In some countries, notably the UK, this coincided with increasing political criticism of the education system for failing to produce the levels of educational achievement required by the 'knowledge economy' and by 'international competition'. Such criticism was closely related to the rise of the 'new public management' in the 1980s, which focused on increasing the 'accountability' of

M. Hammersley (✉)
Open University UK, England, UK
e-mail: martyn.hammersley@open.ac.uk

O. Zawacki-Richter et al. (eds.), *Systematic Reviews in Educational Research*,
https://doi.org/10.1007/978-3-658-27602-7_2

public sector workers, including teachers, through setting targets, highlighting 'best practice', and monitoring performance (Hammersley 2000, 2013; Lane 2000). This was held to be the most effective way of 'driving up standards', and thereby improving national economic performance. In this context, it was complained not just that there was insufficient educational research of high quality relevant to key educational issues (Hargreaves 1996/2007; see also Hammersley 1997a), but also that the findings available had not been synthesised systematically so as to provide the practical guidance required. In an attempt to remedy this, not only were funds directed into increasing the amount of policy- and practice-relevant research on teaching and learning, but also into producing systematic reviews of findings relating to a wide range of educational issues (Davies 2000; Oakley et al. 2005).

In the context of medicine, systematic reviewing was usually conceived as summarising results from RCTs, via meta-analysis; and, as already noted, such trials were often regarded as the gold standard for investigations designed to determine the effectiveness of any kind of 'treatment'. However, in the 1990s relatively few RCTs had been carried out in education and therefore many of the systematic reviews produced had to rely on evidence from a wider range of research methods. One effect of this was to encourage the use of alternative ways of synthesising research findings, including ones that could be applied to findings from qualitative studies (see Barnett-Page and Thomas 2009; Dixon-Woods et al. 2005; Hammersley 2013, Chap. 11; Hannes and Macaitis 2012; Pope et al. 2007; Thomas et al. 2017). Furthermore, qualitative research began to be seen as providing a useful supplement to quantitative findings: it was believed that, while the latter indicated whether a policy or practice is effective in principle, these other kinds of evidence could offer useful contextual information, including about how the policy or practice is perceived and responded to by the people involved, which could moderate judgments about its likely effectiveness 'in the real world'. Subsequently, along with a shift towards giving a role to representatives of potential users of reviews in designing them, there was also recognition that some aspects of systematic reviews are not appropriate in relation to qualitative research, so that there came to be recognition of the need for 'integrative reviews' (Victor 2008) or 'configurative reviews' (Gough et al. 2013) as a variant of or complement to them.

Of course, 'systematic' is a laudatory label, so anything that is not systematic would generally be regarded as inadequate. Indeed, advocacy of systematic reviews often involved sharp criticism of 'traditional' or 'narrative' reviews, these being dismissed as "subjective" (Cooper 1998, p. xi), as involving "haphazard" (Slavin 1986, p. 6) or "arbitrary" (p. 10) selection procedures, as frequently summarising "highly unrepresentative samples of studies in an unsystematic and

uncritical fashion" (Petticrew and Roberts 2006, p. 5), or (even more colourfully) as amounting to "selective, opinionated and discursive rampages through the literature which the reviewer happens to know about or can easily lay his or her hands on" (Oakley 2001/2007, p. 96).[1] Given this, it is perhaps not surprising that the concept of systematic review was itself subjected to criticism by many social scientists, for example being treated as reflecting an outdated positivism (Hammersley 2013, Chap. 8; MacLure 2005; Torrance 2004). And discussions between the two sides often generated more heat than light (see, for instance, Chalmers 2003, 2005; Hammersley 2005, 2008a; Oakley 2006).

There are several problems involved in evaluating the methodological arguments for and against systematic reviews. These include the fact that, as just noted, the concept of systematic review became implicated in debates over qualitative versus quantitative method, and the philosophical assumptions these involve. Furthermore, like any other research strategy, systematic reviewing can have disadvantages, or associated dangers, as well as benefits. Equally important, it is an ensemble of components, and it is possible to accept the value of some of these without accepting the value of all. Finally, reviews can serve different functions and what is good for one may be less so for others.[2]

2 Criticism of Systematic Reviews

Because systematic review was associated with the evidence-based practice movement, the debates around it were closely linked with wider social and political issues. For instance, the idea that medical decisions should be determined by the results of clinical trials was challenged (not least, by advocates of 'personalised medicine'), and there was even more reaction in other fields against the notion that good professional practice is a matter of 'implementing' proven 'treatments', as against exercising professional expertise to evaluate what would be best in particular circumstances. As Torrance (2004) remarks: "Systematic reviewing can thus be seen as part of a larger discourse of distrust, of professionals and of expertise, and the increasing procedurisation of decision-making processes in risk-averse organisations" (p. 3).

[1] At other times, systematic reviewing is presented as simply one form among others, each serving different purposes (see Petticrew and Roberts 2006, p. 10).

[2] For practical guides to the production of systematic reviews, see Petticrew and Roberts (2006) and Gough et al. (2017).

It was also argued that an emphasis on 'what works' obscures the value issues involved in determining what is good policy or practice, often implicitly taking certain values as primary. Arguments about the need for evidence-based, or evidence-informed, practice resulted, it was claimed, in education being treated as the successful acquisition of some institutionally-defined body of knowledge or skill, as measured by examination or test results; whereas critics argued that it ought to be regarded as a much broader process, whether of a cognitive kind (for instance, 'learning to learn' or 'independent learning') or moral/political/religious in character' (learning to understand one's place in the world and act accordingly, to be a 'good citizen', etc.). Sometimes this sort of criticism operated at a more fundamental level, challenging the assumption that teaching is an instrumental activity (see Elliott 2004, p. 170–176). The argument was, instead, that it is a process in which values, including cognitive learning, are realised intrinsically: that they are internal goods rather than external goals. Along these lines, it was claimed that educational research of the kind assumed by systematic reviewing tends necessarily to focus on the acquisition of superficial learning, since this is what is easily measurable. In this respect systematic reviews, along with the evidence-based practice movement more generally, were criticised for helping to promote a misconceived form of education, or indeed as anti-educational.

There was also opposition to the idea, implicit in much criticism of educational research at the time when systematic reviewing was being promoted, that the primary task of this research is to evaluate the effectiveness of policies and practices. Some insisted that the main function of social and educational research is socio-political critique, while others defended a more academic conception of research on educational institutions and practices. Here, too, discussion of systematic reviewing became caught up in wider debates, this time about the proper functions of social research and, more broadly, about the role of universities in society.

While this broader background is relevant, I will focus here primarily on the specific criticisms made of systematic reviewing. These tended to come from two main sources: as already noted, one was qualitative researchers; the other was advocates of realist evaluation and synthesis (Pawson et al. 2004; Pawson 2006b). Realists argued that what is essential in evaluating any policy or practice is to identify the causal mechanism on which it is assumed to rely, and to determine whether this mechanism actually operates in the world, and if so under what conditions. Given this, the task of reviewing is not to find all relevant literature about the effects of some policy, but rather to search for studies that illuminate the causal processes assumed to be involved (Pawson 2006a; Wong 2018). Furthermore, what is important, often, is not so much the validity of the evidence but

its fruitfulness in generating and developing theoretical ideas about causal mechanisms. Indeed, while realists recognise that the validity of evidence is important when it comes to testing theories, they emphasise the partial and fallible character of all evidence, and that the search for effective causal mechanisms is an ongoing process that must take account of variation in context, since some differences in context can be crucial for whether or not a causal mechanism operates and for what it produces. As a result, realists do not recommend exhaustive searches for relevant material, or the adoption of a fixed hierarchy of evidential quality. Nor are they primarily concerned with aggregating findings, but rather with using these to develop and test hypotheses deriving from theories about particular types of policy-relevant causal process. What we have here is a completely different conception of what the purpose of reviews is from that built into most systematic reviewing.

Criticism of systematic reviewing by qualitative researchers took a rather different form. It was two-pronged. First, it was argued that systematic reviewing downplays the value of qualitative research, since the latter cannot supply what meta-analysis requires: measurements providing estimates of effect sizes. As a result, at best, it was argued, qualitative findings tend to be accorded a subordinate role in systematic reviews. A second line of criticism concerned what was taken to be the positivistic character of this type of review. One aspect of this was the demand that systematic reviewers must employ *explicit procedures* in selecting and evaluating studies. The implication is that reviews must not rely on current judgments by researchers in the field about what are the key studies, or about what is well-established knowledge; nor must they depend upon reviewers' own background expertise and judgment. Rather, a technical procedure is to be employed—one that is held to provide 'objective' evidence about the current state of knowledge. It was noted that this reflects a commitment to procedural objectivity (Newell 1986; Eisner 1992), characteristic of positivism, which assumes that subjectivity is a source of bias, and that its role can and must be minimised. Generally speaking, qualitative researchers have rejected this notion of objectivity. The contrast in orientation is perhaps most clearly indicated in the advocacy, by some, of 'interpretive reviews' (for instance, Eisenhart 1998; see Hammersley 2013, Chap. 10).

In the remainder of this chapter, I will review the distinctive methodological features of systematic reviews and evaluate them in light of these sources of criticism. I take these features to be: exhaustive searching for relevant literature; explicit selection criteria regarding relevance and validity; and synthesis of relevant findings.

3 Exhaustive Searching for Relevant Material

One of the criticisms that advocates of systematic review directed at traditional reviews was that they were selective in their identification of relevant literature, rather than being the product of an exhaustive search. They argued not just that, as a result, some relevant literature was not taken into account, but also that this selectivity introduced bias, analogous to sampling bias. This argument relies on a parallel being drawn with social surveys of people (see Petticrew and Roberts 2006, p. 15; Shadish 2006, p. vii).

There is not, or should not be, any disagreement about the need to make good use of previous studies in summarising existing knowledge. And this clearly requires that an effective search is carried out (see Hart 2001). Furthermore, while there is a danger of comparing systematic review as an ideal type with relatively poor actual examples of traditional reviews,[3] there is certainly a difference between the two types of review in the degree to which the search for relevant literature aims to be exhaustive. It is also probably true that the searches carried out in producing many traditional reviews missed relevant literature. Nevertheless, the demand for *exhaustive* searches is problematic.

A first point is that any simple contrast between exhaustive coverage and a biased sample is misleading, since the parallel with social surveys is open to question. At its simplest, the aim of a systematic review is to determine whether a particular type of treatment produces a particular type of effect, and this is a different enterprise from seeking to estimate the distribution of features within some population. The set of studies identified by an exhaustive search may still be a biased sample of the set of studies that could have been done, which would be the appropriate population according to this statistical line of thinking.[4] Furthermore, pooling the results from all the studies that have been done will not give us sound knowledge unless our judgments about the likely validity of the findings from each study are accurate. Increasing the size of the pool from which studies are selected does not, in itself, guarantee any increase in the likely validity of a review's findings.

[3]There are, inevitably, often failings in how systematic reviews are carried out, even in their own terms (see Petticrew and Roberts 2006, p. 270; Thompson 2015).

[4]Indeed, they may not even be a representative sample of the studies that have actually been done, as a result of publication bias: the tendency for studies that find no relationship between the variables investigated to be much less likely to be published than those that produce positive findings.

There are also practical reasons for questioning the ideal of exhaustive searching. Searching for relevant literature usually reaches a point where the value of what is still to be discovered is likely to be marginal. This is not to deny that, because of the patchiness of literatures, it is possible that material of high relevance may be found late on in a search, or missed entirely. But the point is that any attempt to eliminate this risk is not cost-free. Given finite resources, whatever time and effort are devoted to searching for relevant literature will be unavailable for other aspects of the reviewing process. For example, one criticism of systematic reviewing is that it results in superficial reading of the material found: with reviewers simply scanning for relevance, and 'extracting' the relevant information so as to assess likely validity on the basis of a checklist of criteria (MacLure 2005).[5] By contrast, qualitative researchers emphasise the need for careful reading and assessment, insisting that this is a hermeneutic task.[6] The key point here is that, as in research generally, trade-off decisions must be made regarding the time and resources allocated among the various sub-tasks of reviewing research literatures. So, rather than an insistence on maximising coverage, judgments should be made about what is the most effective allocation of time and energy to the task of searching, as against others.

There are also some questions surrounding the notion of relevance, as this is built into how a search is carried out. Where, as with many systematic reviews, the task is to find literature about the effects of a specific policy or practice, there may be a relatively well-defined boundary around what would count as relevant. By contrast, in reviews serving other functions, such as those designed to summarise the current state of knowledge in a field, this is not always the case. Here, relevance may not be a single dimension: potentially relevant material could extend in multiple directions. Furthermore, it is often far from clear where the limit of relevance lies in any of these directions. The principle of exhaustiveness is hard to apply in such contexts, even as an ideal; though, of course, the need to attain sufficient coverage of relevant literature for the purposes of the review remains. Despite these reservations, the systematic review movement has served a useful general function in giving emphasis to the importance of active searching for relevant literature, rather than relying primarily upon existing knowledge in a field.

[5]For an example of one such checklist, from the health field, see https://www.gla.ac.uk/media/media_64047_en.pdf (last accessed: 20.02.19).

[6]For an account of what is involved in understanding and assessing one particular type of research, see Hammersley (1997b).

4 Transparent Methodological Assessment of Studies

The second key feature of systematic reviewing is that explicit criteria should be adopted, both in determining which studies found in a search are sufficiently relevant to be included, and in assessing the likely validity of research findings. As regards relevance, clarity about how this was determined is surely a virtue in reviews. Furthermore, it is true that many traditional reviews are insufficiently clear not just about how they carried out the search for relevant literature but also about how they determined relevance. At the same time, as I noted, in some kinds of review the boundaries around relevance are complex and hard to determine, so that it may be difficult to give a very clear indication of how relevance was decided. We should also note the pragmatic constraints on providing information about this and other matters in reviews, these probably varying according to audience. As Grice (1989) pointed out in relation to communication generally, the quantity or amount of detail provided must be neither too little *nor too much*. A happy medium as regards how much information about how the review was carried out should be the aim, tailored to audience; especially given that complete 'transparency' is an unattainable ideal.

These points also apply to providing information about how the validity of research findings was assessed for the purposes of a review. But there are additional problems here. These stem partly from pressure to find a relatively quick and 'transparent' means of assessing the validity of findings, resulting in attempts to do this by identifying standard features of studies that can be treated as indicating the validity of the findings. Early on, the focus was on overall research design, and a clear hierarchy was adopted, with RCTs at the top and qualitative studies near the bottom. This was partly because, as noted earlier, qualitative studies do not produce the sort of findings required by systematic reviews; or, at least, those that employ meta-analysis. However, liberalisation of the requirements, and an increasing tendency to treat meta-analysis as only one option for synthesising findings, opened up more scope for qualitative and other non-experimental findings to be included in systematic reviews (see, for instance, Petticrew and Roberts 2006). But the issue of how the validity of these was to be assessed remained. And the tendency has been to insist that what is required is a list of specified design features that must be present if findings are to be treated as valid.

This raised particular problems for qualitative research. There have been multiple attempts to identify criteria for assessing such work that parallel those

generally held to provide a basis for assessing quantitative studies, such as internal and external validity, reliability, construct validity, and so on. But not only has there been some variation in the qualitative criteria produced, there have also been significant challenges to the very idea that assessment depends upon criteria identifying specific features of research studies (see Hammersley 2009; Smith 2004). This is not the place to rehearse the history of debates over this (see Spencer et al. 2003). The key point is that there is little consensus amongst qualitative researchers about how their work should be assessed; indeed, there is considerable variation even in judgments made about particular studies. This clearly poses a significant problem for incorporating qualitative findings into systematic reviews; though there have been attempts to do this (see Petticrew and Roberts 2006, Chap. 6), or even to produce qualitative systematic reviews (Butler et al. 2016), as well as forms of qualitative synthesis some of which parallel meta-analysis in key respects (Dixon-Woods et al. 2005; Hannes and Macaitis 2012; Sandelowski and Barroso 2007).

An underlying problem in this context is that qualitative research does not employ formalised techniques. Qualitative researchers sometimes refer to what may appear to be standard methods, such as 'thick description', 'grounded theorising', 'triangulation', and so on. However, on closer inspection, none of these terms refers to a single, standardised practice, but instead to a range of only broadly defined practices. The lack of formalisation has of course been one of the criticisms made of qualitative research. However, it is important to recognise, first of all, that what is involved here is a difference from quantitative research *in degree*, not a dichotomy. Qualitative research follows loose guidelines, albeit flexibly. And quantitative research rarely involves the mere application of standard techniques: to one degree or another, these techniques have to be adapted to the particular features of the research project concerned.

Moreover, there are good reasons why qualitative research is resistant to formalisation. The most important one is that such research relies on unstructured data, data not allocated to analytic categories at the point of collection, and is aimed at *developing* analytic categories not testing pre-determined hypotheses. It therefore tends to produce sets of categories that fall short of the requirements of mutual exclusivity and exhaustiveness required for calculating the frequencies with which data fall into one category rather than another—which are the requirements that govern many of the standard techniques used by quantitative researchers, aside from those that depend upon measurement. The looser form of categorisation employed by qualitative researchers facilitates the development of analytic ideas, and is often held to capture better the complexity of the social world. Central here is an emphasis on the role of people's interpretations and

actions in producing outcomes in contingent ways, rather than these being produced by deterministic mechanisms. It is argued that causal laws are not available, and therefore, rather than reliable predictions, the best that research can offer is enlightenment about the complex processes involved, in such a manner as to enable practitioners and policymakers themselves to draw conclusions about the situations they face and make decisions about what policies or practices it would be best to adopt. Qualitative researchers have also questioned whether the phenomena of interest in the field of education are open to counting or measurement, for example proposing thick description instead. These ideas have underpinned competing forms of educational evaluation that have long existed (for instance 'illuminative evaluation', 'qualitative evaluation' or 'case study') whose character is sharply at odds with quantitative studies (see, for instance, Parlett and Hamilton 1977). In fact, the problems with RCTs, and quantitative evaluations more generally, had already been highlighted in the late 1960s and early 1970s.

A closely related issue is the methodological diversity of qualitative research in the field of education, as elsewhere: rather than being a single enterprise, its practitioners are sharply divided not just over methods but sometimes in what they see as the very goal or product of their work. While much qualitative inquiry shares with quantitative work the aim of producing sound knowledge in answer to a set of research questions, some qualitative researchers aim at practical or political goals—improving educational practices or challenging (what are seen as) injustices—or at literary or artistic products—such as poetry, fiction, or performances of some sort (Leavy 2018). Clearly, the criteria of assessment relevant to these various enterprises are likely to differ substantially (Hammersley 2008b, 2009).

Aside from these problems specific to qualitative research, there is a more general issue regarding how research reviews can be produced for lay audiences in such a way as to enable them to evaluate and trust the findings. The ideal built into the concept of systematic review is assessment criteria that *anyone* could use successfully to determine the validity of research findings, simply by looking at the research report. However, it is doubtful that this ideal could ever be approximated, even in the case of quantitative research. For example, if a study reports random allocation to treatment and control groups, this does not tell us how successfully randomisation was achieved in practice. Similarly, while it may be reported that there was double blinding, neither participants nor researcher knowing who had been allocated to treatment and control groups, we do not know how effectively this was achieved in practice. Equally significant, neither randomisation nor blinding eliminate all threats to the validity of research

findings. My point is not to argue against the value of these techniques, simply to point out that, even in these relatively straightforward cases, statements by researchers about what methods were used do not give readers all the information needed to make sound assessments of the likely validity of a study's findings. And this problem is compounded when it comes to lay reading of reviews. Assessing the likely validity of the findings of studies and of reviews is necessarily a matter of *judgment* that will rely upon background knowledge—including about the nature of research of the relevant kinds and reviewing processes—that lay audiences may not have, and may be unable or unwilling to acquire. This is true whether the intended users of reviews are children and parents or policymakers and politicians.

That there is a problem about how to convey research findings to lay audiences is undoubtedly true. But systematic reviewing does not solve it. And, as I have indicated, there may be significant costs involved in the attempt to make reviewers' methodological assessment of findings transparent through seeking to specify explicit criteria relating to the use of standardised techniques.

5 Synthesis of Findings

It is important to be clear about exactly what 'synthesis' means, and also to recognise the distinction between the character or purpose of synthesis and the means employed to carry it out. At the most basic level, synthesis involves putting together findings from different studies; and, in this broad sense, many traditional as well as systematic reviews engage in this process, to some degree. However, what is involved in most systematic reviews is a very particular kind of synthesis: the production of a summary measure of the likely effect size of some intervention, based on the estimates produced by the studies reviewed. The assumption is that this is more likely to be accurate than the findings of any of individual studies because the number of cases from which data come is greater. Another significant feature of systematic reviews is that a formal and explicit method is employed, such as meta-analysis. These differences between traditional and systematic reviews raise a couple of issues.

One concerns the assumption that what is to be reviewed is a set of studies aimed at identifying the effects of a 'treatment' of some kind. Much reviewing of literature in the field of education, and in the social sciences more generally, does not deal exclusively with studies of this kind. In short, there are differences between systematic and other kinds of review as regards what is being

synthesised and for what purpose. Traditional reviews often cover a range of types of study, these not only using different methods but also aiming at different types of product. Their findings cannot be added together, but may complement one another in other ways—for example relating to different aspects of some problem, organisation, or institution. Furthermore, the aim, often, is to identify key landmarks in a field, in theoretical and/or methodological terms, or to highlight significant gaps in the literature, or questions to be addressed, rather than to determine the answer to a specific research question. Interestingly, some forms of qualitative synthesis are close to systematic review in purpose and character, while others—such as meta-ethnography—are concerned with theory development (see Noblit and Hare 1988; Toye et al. 2014).

What kind of synthesis or integration is appropriate depends upon the purpose(s) of, and audience(s) for, the particular review. As I have hinted, one of the problems with the notion of systematic reviewing is that it tends to adopt a standard model. It may well be true that for some purposes and audiences the traditional review does not engage in sufficient synthesis of findings, but this is a matter of judgment, as is what kind of synthesis is appropriate. As we saw earlier, realist evaluators argue that meta-analysis, and forms of synthesis modelled on it, may not be the most appropriate method even where the aim is to address lay audiences about what are the most effective policies or practices. They also argue that this kind of synthesis largely fails to answer more specific questions about what works for whom, when, and where—though there is, perhaps, no reason in principle why systematic reviews cannot address these questions. For realists what is required is not the synthesis of findings through a process of aggregation but rather to use previous studies in a process of theory building aimed at identifying the key causal mechanisms operating in the domain with which policymakers or practitioners are concerned. This seems to me to be a reasonable goal, and one that has scientific warrant.

Meanwhile, as noted earlier, some qualitative researchers have adopted an even more radical stance, denying the possibility of useful generalisations about sets of cases. Instead, they argue that inference should be from one (thickly described) case to another, with careful attention to the dimensions of similarity and difference, and the implications of these for what the consequences of different courses of action would be. However, while this is certainly a legitimate form of inference in which we often engage, it seems to me that it involves implicit reliance on ideas about what is likely to be generally true. It is, therefore, no substitute for generalisation.

A second issue concerns, once again, the advantages and disadvantages of standardisation or formalisation.[7] Traditional reviews tend to adopt a less standardised, and often less explicit, approach to synthesis; though the development of qualitative synthesis has involved a move towards more formal specification. Here, as with the methodological assessment of findings, it is important to recognise that exhaustive and fully transparent specification of the reviewing process is an ideal that is hard to realise, since judgment is always involved in the synthesis process. Furthermore, there are disadvantages to pursuing this ideal of formalisation *very* far, since it downgrades the important role of imagination and creativity, as well as of background knowledge and scientific sensibility. Here, as elsewhere, some assessment has to be made about the relative advantages and disadvantages of formalisation, necessarily trading these off against one another, in order to find an appropriate balance. A blanket insistence that 'the more the better', in this area as in others, is not helpful.

6 Conclusion

In this chapter I have outlined some of the main criticisms that have been made of systematic reviews, and looked in more specific terms at issues surrounding their key components: exhaustive searching; the use of explicit criteria to identify relevant studies and to assess the validity of findings; and synthesis of those findings. It is important to recognise just how contentious the promotion of such reviews has been, partly because of the way that this has often been done through excessive criticism of other kinds of review, and because the effect has been seen as downgrading some kinds of research, notably qualitative inquiry, at the expense of others. But systematic reviews have also been criticised because of the assumptions on which they rely, and here the criticism has come not just from qualitative researchers but also from realist evaluators.

It is important not to see these criticisms as grounds for dismissing the value of systematic reviews, even if this is the way they have sometimes been formulated. For instance, most researchers would agree that in any review an adequate search of the literature must be carried out, so that what is relevant is identified as clearly as possible; that the studies should be properly assessed in methodological

[7]For an account of the drive for standardisation, and thereby for formalisation, in the field of health care, and of many of the issues involved, see Timmermans and Berg (2003).

terms; and that this ought to be done, as far as possible, in a manner that is intelligible to readers. They might also agree that many traditional reviews in the past were not well executed. But many would insist, with Torrance (2004, p. 3), that 'perfectly reasonable arguments about the transparency of research reviews and especially criteria for inclusion/exclusion of studies, have been taken to absurd and counterproductive lengths'. Thus, disagreement remains about what constitutes adequate search for relevant literature, how studies should be assessed, what information can and ought to be provided about how a review was carried out, and what degree and kind of synthesis should take place.

The main point I have made is that reviews of research literatures serve a variety of functions and audiences, and that the form they need to take, in order to do this effectively, also varies. While being 'systematic', in the tendentious sense promoted by advocates of systematic reviewing, may serve some functions and audiences well, this will not be true of others. Certainly, any idea that there is a single standard form of review that can serve all purposes and audiences is a misconception. So, too, is any dichotomy, with exhaustiveness and transparency on one side, bias and opacity on the other. Nevertheless, advocacy of systematic reviews has had benefits. Perhaps its most important message, still largely ignored across much of social science, is that findings from single studies are likely to be misleading, and that research knowledge should be communicated to lay audiences via reviews of all the relevant literature. While I agree strongly with this, I demur from the conclusion that these reviews should always be 'systematic'.

References

Barnett-Page, E., & Thomas, J. (2009). Methods for the synthesis of qualitative research: a critical review, *BMC Medical Research Methodology*, 9:59. https://doi.org/10.1186/1471-2288-9-59.

Butler, A., Hall, H., & Copnell, B. (2016). A guide to writing a qualitative systematic review protocol to enhance evidence-based practice in nursing and health care. *Worldviews on Evidence-based Nursing, 13*(3), 241–249. https://doi.org/10.1111/wvn.12134.

Chalmers, I. (2003). Trying to do more good than harm in policy and practice: the role of rigorous, transparent, up-to-date evaluations. *Annals of the American Academy of Political and Social Science*, 589, 22–40. https://doi.org/10.1177/0002716203254762.

Chalmers, I. (2005). If evidence-informed policy works in practice, does it matter if it doesn't work in theory? *Evidence and Policy*, *1*(2), 227–42. https://doi.org/10.1332/1744264053730806.

Cooper. H. (1998). *Research synthesis and meta-analysis*, Thousand Oaks CA: Sage.

Davies, P. (2000). The relevance of systematic reviews to educational policy and practice. *Oxford Review of Education*, 26(3/4), 365–78. https://doi.org/10.1080/713688543.

Dixon-Woods, M., Agarwall, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: a review of possible methods. *Journal of Health Service Research and Policy*, *10*(1), 45–53. https://doi.org/10.1177/135581960501000110.

Eisenhart, M. (1998). On the subject of interpretive reviews. *Review of Educational Research*, *68*(4), 391–399. https://www.jstor.org/stable/1170731.

Eisner, E. (1992). Objectivity in educational research. *Curriculum Inquiry*, *22*(1), 9–15. https://doi.org/10.1080/03626784.1992.11075389.

Elliott, J. (2004). Making evidence-based practice educational. In G. Thomas & R. Pring (Eds.), *Evidence-Based Practice in Education* (pp. 164–186), Maidenhead: Open University Press.

Gough, D., Oliver, S., & Thomas, J. (2013). *Learning from research: systematic reviews for informing policy decisions: A quick guide*. London: Alliance for Useful Knowledge.

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

Grice, P. (1989). *Studies in the way of words*. Cambridge MA: Harvard University Press.

Hammersley, M. (1997a). Educational research and teaching: a response to David Hargreaves' TTA lecture. *British Educational Research Journal*, *23*(2), 141–161. https://doi.org/10.1080/0141192970230203.

Hammersley, M. (1997b). *Reading Ethnographic Research* (2nd edition). London: Longman.

Hammersley, M. (2000). Evidence-based practice in education and the contribution of educational research. In S. Reynolds & E. Trinder (Eds.), *Evidence-based practice* (pp. 163–183), Oxford: Blackwell.

Hammersley, M. (2005). Is the evidence-based practice movement doing more good than harm? *Evidence and Policy*, *1*(1), 1–16. https://doi.org/10.1332/1744264052703203.

Hammersley, M. (2008a). Paradigm war revived? On the diagnosis of resistance to randomised controlled trials and systematic review in education. *International Journal of Research and Method in Education*, *31*(1), 3–10. https://doi.org/10.1080/17437270801919826.

Hammersley, M. (2008b). Troubling criteria: a critical commentary on Furlong and Oancea's framework for assessing educational research. *British Educational Research Journal*, *34*(6), 747–762. https://doi.org/10.1080/01411920802031468.

Hammersley, M. (2009). Challenging relativism: the problem of assessment criteria. *Qualitative Inquiry*, *15*(1), 3–29. https://doi.org/10.1177/1077800408325325.

Hammersley, M. (2013). *The myth of research-based policy and practice*. London: Sage.

Hammersley, M. (2014). Translating research findings into educational policy or practice: the virtues and vices of a metaphor. *Nouveaux c@hiers de la recherche en éducation*, *17*(1), 2014, 54–74, and *Revue suisse des sciences de l'éducation*, *36*(2), 213–228.

Hannes K. & Macaitis K. (2012). A move to more systematic and transparent approaches in qualitative evidence synthesis: update on a review of published papers. *Qualitative Research 12*(4), 402–442. https://doi.org/10.1177/1468794111432992.

Hargreaves, D. H. (2007). Teaching as a research-based profession: Possibilities and prospects', Annual Lecture. In M. Hammersley (Ed.), *Educational research and evidence-based practice* (pp. 3–17). London, Sage. (Original work published 1996)

Hart, C. (2001) *Doing a literature search: A comprehensive guide for the social sciences*. London: Sage.
Lane, J. E. (2000) *New Public Management*. London: Routledge.
Leavy, P. (Ed.) (2018). *Handbook of arts-based research*. New York: Guilford Press.
MacLure, M. (2005). Clarity bordering on stupidity: where's the quality in systematic review? *Journal of Education Policy*, *20*(4), 393–416. https://doi.org/10.1080/02680930500131801.
Newell, R. W. (1986). *Objectivity, empiricism and truth*. London: Routledge and Kegan Paul.
Nisbet, J., & Broadfoot, P. (1980). *The impact of research on policy and practice in education*. Aberdeen: Aberdeen University Press.
Noblit G., & Hare R. (1988). *Meta-ethnography: synthesising qualitative studies.* Newbury Park CA: Sage.
Oakley, A. (2006). Resistances to "new" technologies of evaluation: education research in the UK as a case study. *Evidence and Policy*, *2*(1), 63–87. https://doi.org/10.1332/174426406775249741.
Oakley, A. (2007). Evidence-informed policy and practice: challenges for social science. In M. Hammersley (Ed.), *Educational research and evidence-based practice* (pp. 91–105). London: Sage. (Original work published 2001)
Oakley A., Gough, D., Oliver, S., & Thomas, J. (2005). The politics of evidence and methodology: lessons from the EPPI-Centre. *Evidence and Policy*, *1*(1), 5–31. https://doi.org/10.1332/1744264052703168.
Parlett, M.R., & Hamilton, D. (1977). Evaluation in illumination: a new approach to the study of innovative programmes. In D. Hamilton et al. (Eds.), *Beyond the numbers game* (pp. 6–22). London: Macmillan.
Pawson, R. (2006a). Digging for nuggets: how "bad" research can yield "good" evidence. *International Journal of Social Research Methodology*, *9*(2), 127–42. https://doi.org/10.1080/13645570600595314.
Pawson, R. (2006b). *Evidence-based policy: a realist perspective*. London: Sage.
Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2004). *Realist synthesis: an introduction.* Retrieved from https://www.researchgate.net/publication/228855827_Realist_Synthesis_An_Introduction (last accessed October 17, 2018).
Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford: Blackwell.
Pope, C., Mays, N., & Popay, J. (2007). Synthesising qualitative and quantitative health evidence: a guide to methods. Maidenhead: Open University Press.
Sandelowski M., & Barroso J. (2007). *Handbook for synthesising qualitative research*. New York: Springer.
Shadish, W. (2006). Foreword to Petticrew and Roberts. In M. Petticrew, & H. Roberts, *Systematic reviews in the social sciences: A practical guide* (pp. vi–ix). Oxford: Blackwell.
Slavin, R. (1986). Best-evidence synthesis: an alternative to meta-analysis and traditional reviews'. *Educational Researcher*, 15(9), 5–11.
Smith, J. K. (2004). Learning to live with relativism. In H. Piper, & I. Stronach (Eds.), *Educational research: difference and diversity* (pp. 45–58). Aldershot, UK: Ashgate.
Spencer, L., Richie, J., Lewis, J., & Dillon, L. (2003). Quality in qualitative evaluation: a framework for assessing research evidence. Prepared by the National Centre for Social

Research on behalf of the Cabinet Office. Retrieved from https://www.heacademy.ac.uk/system/files/166_policy_hub_a_quality_framework.pdf (last accessed October 17, 2018).

Thomas, J., O'Mara-Eves, A., Harden, A., & Newman, M. (2017). Synthesis methods for combining and configuring textual and mixed methods data. In D. Gough, S. Oliver, S. & J. Thomas (Eds.), *An introducation to systematic reviews* (2nd edition, pp. 181–210). London: Sage.

Thompson, C. (2015). Sins of omission and commission in systematic reviews in nursing: a commentary on McRae et al. (2015). *International Journal of Nursing Studies*, *52*(7), 1277–1278. https://doi.org/10.1016/j.ijnurstu.2015.03.003.

Timmermans, S., & Berg, M. (2003). *The gold standard: the challenge of evidence-based medicine and standardization in health care*. Philadelphia PA: Temple University Press.

Torrance, H. (2004, June). Systematic reviewing—the 'call centre' version of research synthesis. Time for a more flexible approach'. Invited presentation to Economic and Social Research Council/Research Capacity Building Network seminar on Systematic Reviewing, University of Sheffield, UK. Retrieved from http://www.esri.mmu.ac.uk/respapers/papers-pdf/seminar-systematicreviewing.pdf (last accessed October 17, 2018).

Toye F., Seers K., Allcock N, Briggs, M., Carr, E., & Barker, K. (2014). Meta-ethnography 25 years on: challenges and insights for synthesising a large number of qualitative studies. *BMC Medical Research Methodology* 14:80. https://doi.org/10.1186/1471-2288-14-80.

Victor, L. (2008). Systematic reviewing. *Social Research Update*, *54*, Retrieved from http://sru.soc.surrey.ac.uk/SRU54.pdf (last accessed October 17, 2018).

Wong, G. (2018). Data gathering in realist reviews: looking for needles in haystacks. In N. Emmel, J. Greenhalgh, A. Manzano, M. Monaghan, & S. Dalkin (Eds.) *Doing realist research* (pp. 131–146). London: Sage.