Systematic Reviewing – the ‘call centre’ version of research synthesis.
Time for a more flexible approach.

Harry Torrance,
Education and Social Research Institute (ESRI),
Manchester Metropolitan University

Invited presentation to ESRC/RCBN seminar on Systematic Reviewing, 24 June 2004,
University of Sheffield

Introduction
The general title of this seminar - ‘situating qualitative research in evidence-based
research and systematic review agendas’ – obviously begs questions about whether
we should attempt to do this at all, as well as how we might. A few brief forays into
search engines and websites certainly suggest that qualitative work is largely ignored
at present. ‘Systematic Reviews’ brings up 1,600,000 hits on Google; ‘Qualitative
Research Reviews’ brings up 644,000 hits; ‘Synthesizing Qualitative Research’ brings
up 23,500; still a lot, but a lot less than 1.6M, and many, of course, irrelevant to the
real topic of concern. Perhaps more surprisingly, searching for ‘Systematic Reviews’
on the ESRC/TLRP website brings up ‘129 documents found’, while searching for
‘Qualitative Research Reviews’ brings up ‘no documents found’, as does searching
for ‘Synthesising Qualitative Research’. So the TLRP obviously hasn’t had much
interest in qualitative research reviews up to now, and there might seem to be a prima
facie case for making some progress here. But making progress also depends on
whether we simply take the definitions and practices of systematic reviewing as
given, or whether we need to broaden the definition of what systematic reviewing
could and should involve, before trying to fit qualitative work into such a
methodological straightjacket.

I am going to argue that, no sooner has qualitative research begun to be considered as
important to include in systematic reviewing, and in particular no sooner have
educational sponsors such as the Teacher Training Agency (TTA) begun to take a
significant interest in systematic reviewing and commission large numbers of such
reviews, than the approach is already past its ‘sell by’ date – certainly with respect to
educational research. The argument is based on two factors concerned with the nature
and politics of knowledge creation and utilisation, and the fitness for purpose of the
methodology. I will then go on briefly to describe how such concerns were addressed
in discussions with a sponsor (the Learning and Skills Research Centre, LSRC) to
renegotiate the boundaries, conduct and reporting of a systematic review.

Systematic Reviewing
The case for developing ‘systematic reviewing’ is based on transparency of process
and clear criteria for including and excluding studies from the review. The case
derives from critiques of so-called ‘narrative reviewing’ which, it is claimed, focuses
on summarising findings, in relation to a particular argument, rather than reviewing
the whole field dispassionately and ‘systematically’ so that the reader can be
confident that all relevant prior knowledge in a field has been included and
summarised (Gough & Elbourne 2002, Oakley 2003). As important, systematic
reviewing claims to include, exclude and, if appropriate, weight, different studies
according to whether or not they are methodologically sound and thus the findings valid and reliable. It is claimed that 'narrative reviews' do not do this, focussing only, as noted above, on summarising findings rather than whether or not the findings of particular studies are well warranted. These are important issues and narrative reviewing should certainly address them if necessary, but, at least in my field of research interest - student assessment - the accusations are hugely overstated. General reviews of the field are not uncommon and certainly do not simply follow a 'party line' of only solipsistically citing others of similar mind without regard to the quality of the research reviewed (cf. Crooks 1988, Wood 1991, Black and Wiliam 1998, Pellegrino, Baxter and Glaser 1999).

Broadly speaking, systematic reviewing, following EPPI-Centre methodology (Oakley 2003) involves:

- specifying a single 'answerable' research question;
- identifying 'search' terms;
- conducting a 'systematic' search of electronic databases, plus hand-searching journals, 'grey' literature, etc;
- defining and reporting explicit inclusion/exclusion criteria, including in relation to research methodology and quality;
- taking initial inclusion/exclusion filtering decisions on the basis of an abstract if it is available or the title if is not (and this is often the case);
- decisions taken by more than one reviewer to increase reliability;
- identification of a 'map' of the field comprising articles to be read and indexed, with further exclusion of articles deemed of less relevance and/or low methodological quality after reading;
- final review of only those texts directly relevant to the research question and of high methodological quality.

The problem for the inclusion of qualitative research is that it has been routinely excluded to date because it is deemed, by definition, to be of low methodological quality – i.e. not involving the measurement of variables and calculation of their relationship(s), and not involving randomised control trial experimental design (RCTs). These exclusion criteria seem to derive from the fact that systematic reviewing was first developed within medical research and qualitative evidence simply didn’t fit within the medical natural science model of what constitutes sound research design and appropriate evidence. Subsequently medical researchers have been arguing the case for the inclusion of qualitative evidence (cf. Barbour and

---

1 Incidentally, Crooks (1988) appears in *Review of Educational Research* which has been produced for more than 70 years, while Pellegrino, Baxter & Glaser (1999) is published in the *Review of Research in Education* which has been appearing annually for more than 25 years, usually taking a particular field each year with a specially appointed team of editors producing each edition. Also, ironically, while the field of student assessment has often produced authoritative reviews of research findings, policymakers take little or no notice of them *cf. the current evidence-free zone of National Testing* (Torrance 2003).
Barbour 2003, Dixon-Woods, Fitzpatrick and Roberts 2001) but arguments about prioritising ‘scientific methods’ which aspire to demonstrate ‘cause and effect’ have nevertheless been central to the development of systematic reviewing in social and educational research (Hargreaves 1996, Oakley 2000). The issue of course, is the extent to which researchers think human behaviour can be explained in causal terms, as opposed to understanding the meaning we give to behaviour in context. Many would argue that social research in general, qualitative research in particular, provides us (including policy makers and practitioners) with evidence which allows us to act more intelligently but cannot guide us unambiguously – research evidence cannot take decisions for us.

Taking the knowledge out of knowledge management

Oakley describes the EPPI-Centre as ‘managing rather than generating knowledge’ (2003, p23) and certainly, to return to the more general argument, the overall critique of ‘narrative reviewing’ seems to derive from a suspicion of expert professional judgement and a concern to produce transparent, auditable procedures for the production of disinterested knowledge – to render the building blocks of policy untouched by human hand, or at least ‘expert’ human hand. Thus Oakley (2003) argues that:

It is the ordinary citizen who is potentially most disadvantaged by the lack of an open, systematic basis of evidence…(p.22)

While Gough & Elbourne (2002) go further and claim that:

The EPPI-Centre…and…the explicit process of review…will democratise the use of research knowledge in society…(p. 231)

By implication, the ‘use of research knowledge in society’ is not engaged in democratically at present. By setting up such binary oppositions, Oakley, Gough and Elbourne position previous approaches to research reviewing as in some sense undemocratic or even anti-democratic: an extraordinary insinuation. Certainly the methods of systematic reviewing seem concerned to exclude expert knowledge from the process as much as possible, categorising and coding in immense detail the sorts of decisions on inclusion and exclusion that reviewers should make (see: EPPI training guidelines @ http://eppi.ioe.ac.uk/EPPIReviewer). In this respect systematic reviewing seems to be reflecting more general trends in institutional decision-making and service management towards quality assurance and audit procedures.

Thus just as bank managers and mortgage lenders are no longer trusted to reach face-to-face judgements about the credit worthiness of clients, and their professional knowledge and experience is codified into software programmes to be operated by clerks in call centres, so researchers are no longer trusted to carry out so-called ‘narrative reviews’. 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 such that, according to one article I recently refereed for a journal ‘anyone interested in a review topic, can, with training in the use of systematic review tools, take part…’; no prior knowledge or expertise required, far less informed professional judgement. 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
Fitness for purpose and value for money
Precisely because of this over-procedurisation however, and of much more likely
significance to sponsors and policy-makers, systematic reviews are neither fit-for-
purpose nor represent value-for-money. The EPPI-Centre ‘REEL Reviews’ website
listed 25 completed reviews as of 18/06/04:
(http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?page=/reel/reviews.htm). They vary in
length from 40-50 pp to almost 200pp. Large parts of the reports are devoted to
describing the methods employed, and listing the keywords, data extraction templates,
etc. in lengthy appendices (as systematic reviewing’s interpretation of transparency
demands). More interestingly, commitment to ‘user engagement’ also demands that a
‘feedback’ site is provided for each review. The first fifteen that I checked, including
such potentially important and contentious topics as assessment, gender, early years
provision and inclusion, all stated that ‘there has been no feedback submitted so far’.
No feedback, probably because, one suspects, no one is even reading these reports.
Oakley (2003) estimates the full cost of a systematic review at £75K or more. Yet, in
return, reviews (certainly in education) now routinely conclude that out of 5,000,
10,000, 15,000 initial ‘hits’, only very few studies could be included in the final
review and such studies tell us very little (and certainly very little compared to what
an ‘expert’ in the field might have been able to identify and summarise anyway, had
they been so commissioned). Oakley herself acknowledges

> the relatively low yield of usable studies derived from the searching
process…Only 0.3% of the initial citations were sufficiently relevant to be
reviewed in depth…This experience of having to search haystacks to find
needles is common in systematic reviews…

(Oakley 2003, p.27)

Thus the methodology ends up telling us about what we don’t know, rather than
helping us to summarise what we do know.

Of course one could argue about whether or not this is because of the poor quality of
the original sources reviewed, or because of the narrowness of the methodology
which excludes far too much, or some combination of the two. One of the routine
‘findings’ of systematic reviews in education is that the quality of reviewed studies is
low and that is why so few can be included in the final review. But it is at least
arguable that this is as much an artefact of the methodology as it is a substantive truth.
Articles published in different journals for different audiences and with different
purposes should not automatically be judged of poor quality just because they do not
fit the template of systematic reviewing. Just as relevant an explanation might be that
research questions only take on meaning in context. It is not wholly surprising
therefore, that questions which are of importance to us now, have not necessarily been
asked (or answered by research) previously – if they had, we would hardly need to ask
them again. Whatever is the case, it means that by excluding the vast majority of
potentially relevant literature in a field, systematic reviewing does not achieve what it
purportedly sets out to achieve. Paying more attention to how and why a range of
studies have been included in a ‘narrative review’ would probably strengthen the
genre; but paying so much attention to how and why the vast majority of studies have been excluded simply renders systematic reviews irrelevant to policy-making.

Just as important if not more so, with respect to the public profile of educational research, these apparent ‘findings’ of systematic reviews add to the discourse of the assumed poor quality of educational research. Other evidence-based groups now routinely cite the poor quality of educational research as a cautionary tale for the rest of the social sciences (e.g. Grayson 2002, Boaz & Ashby 2003). They usually quote Hargreaves (1996) and Tooley & Darby (1998) as their key sources (as does Oakley 2003). This is highly ironic of course since neither of these sources is based on what would pass muster as a ‘systematic review’. Hargreaves (1996) was a largely polemical invited lecture and Tooley and Darby’s (1998) report was a highly selective “critique” published by OfSTED, the Chief Inspector of which, at the time, was Chris Woodhead, on record as regarding educational research as “badly written dross” (Woodhead 1998, p. 51). Such obviously biased sources would hardly qualify as ‘grey literature’, let alone peer reviewed empirical research; quite the reverse, they appropriate the legitimacy of research for partisan point-scoring – exactly the sort of literature that one would imagine systematic reviewers would automatically exclude.

The more general issue however, is that by retreating into procedure and technology, systematic reviewing cuts itself off from the actual processes by which social scientific judgements are made and by which knowledge advances – i.e. reflective synthesis within a discipline which combines prior knowledge with new evidence in the development of contingent ‘middle range theories’ (Pawson 2002) which seek both to explain what has happened (evaluation) and guide what might happen in the future (policy). Some argue that qualitative research can be incorporated in systematic reviewing by the use of Bayesian techniques (Gorrard forthcoming; Roberts et. al. 2002) but this displaces the issue rather than resolves it. Bayesian techniques involve transforming qualitative evidence into numerical indicators by the use of experts rank-ordering the importance of key variables discernable from qualitative studies, transforming these into probabilities (of which variables are likely to be most important) and then including them in quantitative meta-analyses. This seems to do little more than add a spurious mathematical accuracy (to three decimal points in Roberts et.al. 2002) to what would be far better left as ‘expert’ judgement. At least we can be appropriately sceptical of expert judgement, precisely because it is usually expressed in narrative form, even if we also regard it as the best available evidence in the circumstances.

Renegotiating a systematic review
It was because of doubts such as these about the efficacy and utility of systematic reviewing that I engaged in negotiations with a sponsor (LSRC) to renegotiate an initial tender for a systematic review. The topic was to do with the impact of assessment on motivation to learn in post-16 education.

The exact review question, included in the original tender document, was:

Do summative assessment and testing have a positive or negative effect on post-16 learners’ motivation for learning in the Learning and Skills Sector?

My response to the tender document was that in my (experienced, expert) judgement, my knowledge of the assessment field in the UK, and indeed internationally,
suggested that there might be relatively little empirical research pertaining directly to the review question, and that a more flexible approach to identifying and reviewing relevant literature might be more useful. The sponsors agreed and a more flexible approach was negotiated.

Before describing what the review did consist of, and which elements of a systematic approach were found useful, it is worthwhile reflecting on why the sponsor might have trusted my judgement in this case. To begin with, I had been previously included in an ‘expert seminar’ by the sponsor when they were consulting with the academic community about key issues for future research, and had met the project manager in this context. Thus I was already known to the sponsor as someone whose judgement could, at least in principle, be trusted. Additionally the response to their original tender had apparently been disappointing, with no-one else offering to do much more than carry out their instructions – i.e. offer a technical service which could probably be carried out in-house just as easily. An ‘expert’ offered (presumably) a wealth of prior knowledge which could be brought to bear on the task, and thus might be said to ‘add value’ to the original brief. Clearly therefore, ongoing contacts between sponsors and academics can bring benefits to the sponsor in terms of trustworthiness and value-for-money (i.e. prior knowledge which does not have to be paid for); and benefits to academics in terms of shaping tenders to suit ongoing research interest and priorities.

With respect to the review process itself, I worked with a research fellow and we borrowed aspects of the systematic approach with respect to defining search terms, articulating inclusion and exclusion criteria, and making use of electronic databases as well as hand-searching journals and consulting users and other experts on key publications in the field. This approach certainly confirmed our initial suspicion that we would find very little research evidence that was directly relevant to the review question. However, employed flexibly, it also allowed us to discover a reasonable amount of what might be termed partially relevant material and which thus allowed us to describe and review the general field of ‘Assessment in post-compulsory education in the UK’. We didn’t directly answer the review question, but we did (we think) provide a focussed and useful report to inform policy (Torrance & Coultas 2004). The report runs to 47 pages, 11 of which are appendices, including 7 which list all the (105) sources reviewed. It is written in what might be termed traditional ‘narrative’ style, tracing themes and issues across different literatures, and highlighting in particular the hitherto rather hermetically sealed research worlds of 16-19 (FE) education, work-based learning and assessment, and informal adult and community education.

In the end, out of 6491 ‘hits’, of which 751 were considered by title to be potentially relevant, 105 papers, articles and research reports were read, summarised and referred to in the report. In this respect it might be said that we adapted and modified systematic review methodology. Rather than regard the 105 titles as constituting the 'mapped' field and then excluding further articles from the final review after reading, we have included them all, in effect reviewing the whole map rather than the limited number of sources that would have survived another round of culling. The

---

2 I am of course very grateful to the LSRC for responding to such a proposal and funding a different sort of review. In particular, the project manager at LSRC, Maggie Greenwood, was very helpful and supportive of the work.
justification is that such a comprehensive summary will significantly improve the utility of the review.

A good deal of the literature in the field encompasses policy proposals and critiques, short papers reporting early or limited empirical findings, and such like. A systematic review would have excluded such sources, probably concluding, *en passant*, that the field was poorly served by research. We however, regard this as an important finding for policy, and (to echo some of the arguments above) not surprising given the generally ‘Cinderella’ status of post-compulsory education in the UK. Some such policy discussions, conference presentations, etc. are therefore included in the review of the field, though these are duly noted as non-empirical, limited empirical, etc. as and when included in the substantive text. The danger of including material which a strictly 'systematic review' would have excluded, is that the empirical basis for any conclusions is more open to challenge. The danger of excluding such material is that the substantive element of the report would have been very short indeed. What we have endeavoured to produce is an overview of research findings and commonly argued positions in the field of ‘assessment in post-compulsory education in the UK’, so that we can distil such lessons as can be learned for policy, but also so that we can identify key areas for further enquiry. In many respects therefore, the conclusions from the review represent issues to be explored and hypotheses to be tested, as much as findings to be disseminated.

So some of the methods of systematic reviewing were employed, but within a context informed by substantial prior knowledge, including knowledge of the theory and practice of assessment, and how assessment interacts with learning, which very much informed our judgements about the focus and organisation of the review. This in turn was combined with sponsor experience and professional judgement from the general field of enquiry (i.e. post-compulsory education) to produce a ‘research review’ which, we would claim is a comprehensive narrative review and does indeed summarise what we can say about the topic rather than what we can’t. Thus knowledge of two research fields were combined (assessment and post-compulsory education) and interrogated in the context of a particular policy interest. The final report is more speculative in its conclusions and more discursive in its presentation than a ‘systematic review’ might promise. But we would argue that this more accurately and appropriately reflects the state of knowledge in the field and the uncertainties that require further investigation and attention from policy.
References

Torrance H. & Coultas J. (2004) Do summative assessment and testing have a positive or negative effect on post-16 learners’ motivation for learning in the learning and skills sector, London, LSRC