'Clarity bordering on stupidity': where's the quality in systematic review?

Maggie MacLure

Paper presented to the British Educational Research Association Annual Conference, Manchester, September 2004

> Clarity bordering on stupidity, a dog's life
> (André Breton)
>
> [C]lever people observe more things and more curiously, but they interpret them … We need a man either very honest, or so simple that he has not the stuff to build up false inventions and give them plausibility; and wedded to no theory. (Montaigne)[1]

Abstract

The article presents a critique of the discourse of "systematic review' in education, as developed and promoted by the EPPI-Centre at the University of London. Based on a close reading of the exhortatory and instructional literature and 30 published reviews, I argue that the approach degrades the status of reading and writing as scholarly activities, replacing these with an alternative lexicon of putatively 'auditable' acts such as screening, mapping, data extraction, synthesis etc. I show how the approach allows and approves research syntheses based on extremely small numbers of primary studies, which seriously compromise their capacity to inform policy or practice. The claims that are made for the transparency and trustworthiness of systematic review do not, I suggest, stand up to scrutiny. The article concludes that systematic review is animated, not just by a dissatisfaction with the unreliability of educational research and researchers (a dissatisfaction characteristic of the 'evidence informed movement' with which it is aligned), but by a fear of language itself.

Introduction

This paper takes a close, and highly critical, look at "systematic review" in education, in the form developed and promoted by the EPPI-Centre at the University of London Institute of Education.[2] Systematic review is a form of research synthesis designed to support evidence-informed policy and practice. It is endorsed by the Department for Education and Skills and other government departments, and by the Teacher Training Agency. The method requires teams of reviewers to search the literature in ever-decreasing circles until they have isolated all and only those studies that address a pre-

defined question, and are also of high enough quality in terms of research design. 'Quality assurance' procedures make sure that all reviewers think, or at least code, alike.[3]

Systematic review is part of the wider evidence-based, or evidence–informed 'movement' (Oakley, 2003: 22). It has been associated with the 'audit culture' and international trends towards control and accountability in knowledge production and use.[4] Its advocates have lent their voices to the criticisms of educational research which have permeated policy and media discourses over the last decade. Exasperated by the inability of educational research to deliver the kind of seemingly hard evidence offered by health and medicine, systematic review favours quantitative methods and embodies a scarcely-concealed positivism that places qualitative research far down the 'credibility hierarchy' (Hammersley, 2001: 545).

My interest is in the *discourse* of systematic review, and the research realities and identities that it invokes. I propose to take apart the details of the inspirational and instructional literature of systematic review, and also to look 'inside' the 30 education reviews published on the EPPI-Centre website in September 2004.

Systematic review, I suggest below, is a backward-looking business. It construes research knowledge as static, transparent and compliant with disciplinary boundaries. It assumes that evidence can be extracted intact from the texts in which it is embedded, and 'synthesised' in a form that is impervious to ambiguities of context, readers' interpretations or writers' arguments (ie bias). Most significantly of all, systematic review systematically degrades the central acts of reviewing: namely, *reading* and *writing,* and the unreliable intellectual acts that these support, such as interpretation, argument and analysis. By replacing reading and writing with an alternate lexicon of scanning, screening, mapping, data-extraction and synthesis, systematic review tries to transform reading and writing into accountable acts. It tries to force their clandestine operations – the bits that happen inside people's heads, or in the incorporeal gaps between decoding and comprehension, thought and expression - up into plain view, where they can be observed, quality-controlled and stripped of interpretation or rhetoric. The assumption appears to be that evidence, once it has been filtered out of the source texts and checked for quality, should be able to speak for itself.

I argue that systematic review, and the discredited view of reading that it embodies, is one of the many fantasies of 'presence' (Derrida, 1988: 236) that animate contemporary education policy in its rage for clarity, transparency and certainty of outcomes. By trying to regulate reading, writing and interpretation, systematic review suppresses aspects of quality in research and scholarship that are at least as important as clarity, countability and accountability – such as intertextual connectivity, critique, interest, expertise, independence, tacit knowledge, chance encounters with new ideas, and dialogic interactions between researcher, 'literature' and 'data'. The tiny dead bodies of knowledge disinterred by systematic review hold little power to generate new understandings, and are more likely, I suggest, to incapacitate researchers than to contribute to research 'capacity'.

The language of systematic review

Let's begin with the general rationale. The language used to describe and justify systematic review is a mix of old-style scientific positivism (systematicity, reliability, rigour, replicability) and the now-familiar rhetoric of the 'audit culture' (transparency, quality assurance, standards).[5] You can see this mix in Oakley's description of the approach as 'explicit, transparent, replicable, accountable and (potentially) updateable (2003: 23). Here is an expanded version of the 'key features' of systematic review:

> an explicit research question to be addressed; transparency of methods used for searching for studies; exhaustive searches which look for unpublished as well as published studies; clear criteria for assessing the quality of studies (both qualitative and quantitative); clear criteria for including or excluding studies based on the scope of the review and quality assessment; joint reviewing to reduce bias; a clear statement of the findings of the review.
> (Evans & Benenfield, 2001: 529)

Some favoured words belong comfortably to both the scientific and the audit-culture discourses, and can invoke these simultaneously. *Clarity* is one example, as in the quote above, which contains three instances of 'clear', plus one of its near relation, 'transparency', and a second-cousin, 'explicit'. The demand for clarity appeals both to scientific objectivity and to accountability, insinuating that other reviewing practices are both unscientific and shady. Calls for clarity, as several critics have noted, are never innocent. They are usually issued by powerful discourse communities acting as if their world view is 'transcendental': that is, existing 'outside history, language [and] context', while everyone else's is partial and limited. (Scheurich, 2000: 344). There is, according to Giroux (1992), a 'politics of clarity' which attempts to regulate the diversity of practices of less powerful communities, by obliging them to render themselves intelligible according to terms set by the status quo.

Another example of double-faced vocabulary is *trustworthiness,* which is demanded both of the research evidence, and of reviewers' assessments of it. As a criterion applied to the evidence, the calculation of 'trustworthiness' asserts positivism's concern with certainty – ie with reliability, validity, rigour and replicability, for which 'trustworthiness' has become an all-purpose synonym. Simultaneously however, the requirement for trustworthiness also does audit work, as a form of discipline of potentially *un*trustworthy academics through 'quality assurance' mechanisms, in the interests of a putatively more democratic knowledge economy. The term carries an unavoidable moral imputation. Even the evidence itself can, it seems, have bad intentions. The 'EPPI-Centre Review Companion' counsels reviewers to check carefully at the 'keywording' stage as 'this can save a considerable amount of time and effort by excluding studies *which have crept through the inclusion/exclusion stage by mistake'* (EPPI-Centre, 2003: p13 of 20; my emphasis). Reviewers are a particular risk in terms of trustworthiness and must be watched carefully. Systematic review 'allows readers to decide for themselves whether the reviewers have looked carefully enough to be able to say they have identified as many

as possible of the studies that could help answer the review's research question.'[6] I will argue below that this is an empty claim, since systematic reviews cannot deliver on the transparency that they promise; and in any case, are written in a standardised format that renders large sections of them virtually unreadable. However in terms of the EPPI-Centre's rhetoric, reviewers and researchers must make themselves transparent to scrutiny, not just by professionals but by the public at large, since 'it is the ordinary citizen who is potentially most disadvantaged by the lack of an open, systematic base of evidence concerning the many interventions that intrude into every corner of their lives' (Oakley, 2003: 22).

Systematic review, and the 'evidence 'movement' (Oakley 2003: 23) from which it emerged, thus continually recycles a 'discourse of distrust' of education professionals and academics (Torrance, 2004: 3). Research, and researchers, are repeatedly reported or implied to be careless, undemocratic, furtive (ie prone to 'hide failures'), biased, incompetent, 'chaotic', 'inward-looking and self-seeking', 'methodologically impoverished' and even potentially life-threatening.[7] All of these accusations can be found, for instance, in Oakley's 2003 article. Most of them are recycled from the small corpus of critiques of educational research that were issued by Hargreaves, Tooley, Hillage and their co-authors, in the mid-to-late 1990s. [8] These originary critiques are repeatedly re-invoked as the root rationale for the necessary rigours of systematic review - despite the fact that none of them are based on evidence that would come close to meeting systematic review's own 'trustworthiness' criteria (see Avis, 2003; Torrance, 2004). Still, the slurs are, in turn, routinely re-circulated by Oakley and other advocates, and not infrequently by the systematic reviewers themselves.

The systematic reviewers are – or at least one would assume them to be – in a difficult position. They are the vanguard army of the rigorous, accountable, democratic, evidence-informed education 'movement'. Many of the review team leaders are acknowledged experts in their field. But from the EPPI perspective they are also members of the abject community of university researchers who need to be kept under scrutiny. To be seen to be trustworthy in the face of such a dense background of calumny, reviewers must submit to extraordinary degrees of monitoring and regulation. The must show all their workings, leaving a 'clear audit trail from primary research to conclusions' (Oakley 2003: 23). They must follow the EPPI guidelines for teaming up with users and writing protocols, employ the EPPI 'tools' for searching databases, 'keywording', 'data extracting' and 'synthesising', and use 'the EPPI-Centre report structure' for writing up. They must check in at every stage of these proceedings with the EPPI-Centre; work to agreed deadlines, and submit their searching, keywording and data-extracting decisions to 'quality assurance' (ie moderation) by EPPI-Centre staff.[9]

Systematic reviewers are intensively managed and monitored throughout the process then, to ensure that they follow procedures to the letter (and I use this cliché intentionally: see below). They are required to embrace not only the positivist epistemological assumptions about the nature of knowledge and research that are built into the EPPI-apparatus, but also to embrace their own subjection to the Taylorist schedules and managerial discipline of the EPPI-Centre regime. Generally, review authors seem to do

this quite readily. One team thanks the EPPI-Centre workers for their 'vigilant support' (Penn et al, 2004: i), and concludes with this testimonial:

> This review, its level of scrutiny and use of evidence, has been a wake-up call. We have been forced into thinking more carefully about the nature of research in early years and the uses it has been put to in justifying policy-making. Even if our own results were not as conclusive as we had hoped, clarifying the issues and highlighting the gaps has been an essential step (Penn et al, 2004: 43).

Without exception, the review authors always have something positive to say about the EPPI experience, though not all are as openly enthusiastic as the authors above. In fact, between the lines of endorsement and apparent compliance one can detect traces in some reviews of a less straightforward accommodation to the strictures of systematic review.[10] Nevertheless, all the reviews devote many, many pages of text to documenting their compliance with the internal and external quality assurance procedures, and making 'transparent' their coding and synthesising decisions.

The corpus of published reviews testifies therefore to the considerable pains that the academic reviewers took to subject themselves to the EPPI-Centre's requirements. And s*ubjection* is entirely the right word here. What we are seeing in operation is a technology for forming new academic subjects. New academic identities are being created in which values such as academic independence, intellectual curiosity and expert judgement are being replaced by industriousness, rule-following, compliance and self-imposed endorsement of 'the hegemonic position of managers' (Avis, 2003: 373). I am not arguing that those older values are necessarily benign, nor indeed that they have ever been pursued all that energetically. For the moment I want simply to point out the conceptual and ontological distance between them and the new academic virtues of hard work and humility. There is nothing novel in this observation. The trend has been widely noted by critics of the audit culture and the evidence movement (cf Strathern, 2000; Avis, 2003). What I want to emphasise particularly here is the way in which compliance and the reconstitution of subjectivity are being enforced *through language itself.*

The tyranny of terminology

Like many other national education policies – such as Ofsted inspection, the National Literacy Strategy, the National Curriculum, and the associated training programmes for all of these - systematic review is a regimen that disciplines subjects (reviewers, researchers, teachers, teacher educators) through language. Subjects are not just required to show that they have followed the quality assurance or coding instructions, but to do so *in the correct language.* Characteristically, systematic review is a fabrication of neologisms, borrowings and buzz-words – data capture and data extraction, synthesis, key-wording, mapping, weighting, inclusion and exclusion criteria, screening, handsearching.[11] The published reviews are remarkably uniform in their use of this vocabulary.

It's not just a matter of vocabulary. Like any language, systematic review imposes structure, levels and taxonomies on the flux of experience. It institutes orders of importance amongst entities and concepts, and establishes which ones 'belong' together. Like other discourses, it sets limits on the ways that the world can be viewed and construed, and establishes what will count as truth. As we have seen, it defines what kind of (academic) subject it is possible to be, and of what matters such subjects can legitimately speak. Thus as texts, systematic reviews not only use the authenticated words, but also rigorously observe the same sequential structure, using ready-made chapter headings and sub-headings. They take on board (though not always without demur – see below), the pre-eminence of scientific certainty, together with that 'credibility hierarchy' of methodological trustworthiness that installs RCTs at the top and qualitative research at the bottom (Hammersley, 2001). They all employ the EPPI machinery for ensuring that they are adopting the same stance towards the as-yet-uncharted field that they are going to review, and for ensuring that they all 'map' that field using the same techniques.

Those who speak the language of systematic review necessarily bend themselves therefore into the new shapes afforded by its disciplinary syntax and its hierarchies of significance. In Kristeva's terms, this is the symbolic order of the Law at work – 'the legal, phallic, linguistic symbolic establishment' (1982:72). It's a regime of truth, no doubt about it (cf Avis, 2003: 373). Does any of this matter? It might be argued, legitimately, that we are always subjected to one truth regime or another - summoned to speak and compelled to be silent, bent into shape by prevailing institutional and cultural conventions (cf Foucault 1980). But this still leaves the question of precisely what shapes and realities are afforded by systematic review. Moreover, while discourses generally exceed the grasp and the control of any one person or group, systematic review is, I would argue, part of a more deliberate policy to reconstitute the cultural and cognitive practices of education professionals *at the very point where words issue from their mouths or their finger-tips.*

It's instructive to look at the silences in the discourse – at what systematic review leaves unspoken. This, after all, is one of the big effects that discourses accomplish: they not only legitimate what can be said, but they also render other things unspeakable. What is *not-said* alongside the much-repeated lexicon of screening, scanning, handsearching, weighting, keywording, data extracting, synthesising and so on? For a start, some of the words that one might associate with the cognitive or intellectual acts of research or reviewing are almost wholly absent. *Analysis* does not appear in the EPPI-Centre Review Companion, for instance. *Interpretation* occurs there only once, in a context where the concern is for its 'consistency'. Perhaps most curiously of all, in a document that is a guide to reviewing, there are only two instances of *read/reading* (both in the section entitled 'Data Extraction').

As Fairclough (1989) points out, 'wording' – the summoning of realities into words – is never an innocent matter. So what kind of difference does it make when significant cognitive and textual acts such as reading, analysing and interpreting are diminished or disappeared? How is reviewing differently real-ised by the insistent 'overwording' of

those alternative terms - scanning, screening and synthesising; mapping, capturing and extracting? This bunch of words carries interesting traces of old and new technologies: nineteenth century industry, exploration and imperialism (all those tools; all that mapping, mining and capturing), mixed up with computing, medical diagnostics and expert systems (scanning, screening, mapping again). Oakley's description of the two main stages of systematic review exhibits this technicist wording:

> a mapping stage, in which relevant literature is captured and systematically keyworded… and an in-depth review stage, in which a subset of the literature … is examined and interrogated in more detail and data extracted from primary studies. (Oakley, 2003: 24; original emphases removed)

The wording of systematic review implies a view of the 'literature' as a bounded territory, containing knowledge that is stable and quantifiable. Reading (or what has replaced it) amounts to the application of (rather aggressive) procedures for mastering the territory and extracting its nuggets of knowledge.

Equally importantly, this new lexicon invokes *auditable* acts. Reviewing is reconstructed as a sequence of tasks that will leave visible and countable residues of themselves, in the form of codes, data entries, maps, weightings and quality assurances. This is how that 'clear audit trail from primary research to conclusions' is supposed to be forged. Acts such as keywording, mapping, data-extracting or synthesising stand in marked contrast to those annoyingly clandestine acts of reading, analysing and interpreting, which happen inside people's heads, are less easy to break down into auditable stages, and are prone to contextual or individual variation, complexity, indirection and qualification.

The lexicon of systematic review thus denies the *textuality* of research; or sees textuality as a kind of enemy or obstacle to be overcome. Evidence is conceptualised as packages of stuff that can be extracted from the base material of texts, weighed and aggregated (synthesised). In bypassing reading and interpretation, which carry the indissoluble marks of their engagement with texts, and replacing analysis and writing with the simple arithmetic of synthesis, the EPPI-lexicon is a fantasy of a text-free knowledge economy, where nuggets of evidence can be extracted from the rhetorical contaminations of persuasion, argument, justification, context, and partiality that are inherent in all texts. This is a dumb, though ancient and persistent delusion.[12]

Reluctant reading

Still, isn't this all 'just' terminological? One might assume that the reviewers must be reading the articles, even if they're now calling it something different. But reading is strangely absent from systematic review in other ways too. For a start, hardly any of the studies identified as potentially relevant at the outset of the review process are read *at all.* In fact, you could describe the EPPI-Centre method as an algorithm for not-reading as much of the research literature as possible. Here's how it works.

First, a 'protocol' is drawn up which formulates the question to be answered by the review, and the inclusion/exclusion criteria to be used to identify relevant research studies. Reviews are expected to stick to the protocol wherever possible, and diversions into unanticipated areas are not encouraged (although the initial question can be tightened or adjusted later in the process). Learning from adjacent areas is not recommended either. So if, for instance, your review question concerns the effects of small-group discussions in science teaching with students aged 11-18, you may decide not to allow yourself to learn from research into group discussions in other school subjects, nor from research involving younger or older learners (cf Bennet et al, 2004: 76). Reviewers are also asked to state in advance how many studies they will be able to 'data extract' within the time and money allocated, so there is little latitude to expand the scope of the review at later stages.[13] Systematic reviewers often set out to map fairly small fields with secure fences, and do not expect to look over the hedge.[14]

Electronic and 'handsearches' are then done to find research studies that are potentially relevant to the review question. The resulting titles and abstracts – *if* the latter are available - are screened to see if they meet the inclusion/exclusion criteria. In most cases, a huge proportion of studies are screened out at this stage for failing to meet the inclusion criteria, which means that most of the potential field is discarded *without the reviewers having read a single study in its entirety*.[15] In many cases, since abstracts are by no means universal, the reviewers will have eliminated a proportion of studies on the basis of their titles alone. Interestingly, despite the concern with transparency, figures seldom seem to be given for this.

Most potential studies are judged ineligible for inclusion, therefore, without being read at all. Like credit card offers in reverse, they are pre-*dis*approved. The next step in the process is to get hold of the full text of papers in the remaining corpus, for keywording, in order to draw up the 'map' of the field. This stage seems inevitably to allow yet more studies to be excluded, when the full text makes it clear that they do not meet the inclusion criteria.

This is not the end of the 'filtering' process. However let us pause to note another sense in which not-reading is recommended at this stage. Screening and keywording does not, it seems, require actually reading the full papers in any detail. The EPPI-Centre Review Companion tells reviewers that (by contrast with the lengthier business of data extraction) 'keywords might be applied to a study in as little as 10 minutes' (p 15/20). Not much reading anticipated at this point therefore.

The next stage usually involves rendering down the corpus into a final selection for full 'data extraction' and 'in-depth review'.[16] This is the point at which something closer to reading in its vernacular sense presumably gets done, since the Review Companion warns that 'the data extraction process takes about 3 hours per study per reviewer', or possibly longer (15/20). However on the evidence of the reviews so far published, pitifully few primary studies make it through to in-depth review, as this is where the 'weighting' of the evidence comes into play. Potential studies are rated 'high', 'medium' or 'low' for various kinds of trustworthiness including research design. As noted above, the

weightings strongly favour controlled and experimental designs with large samples. Since, as is widely known, very little educational research is of this kind, it is presumably not surprising to the reviewers and sponsors that further big reductions in the corpus of relevant studies are often in order at this last stage.

The evidential base of many of the in-depth reviews is unimpressive. Out of 28 reports on the EPPI-Centre website in September 2004, 16 were based on in-depth reviews of fewer than 10 primary studies (including 3 that were based on only 2 primary studies).[17] A list of all the published reports, arranged by size of their in-depth reviews, can be found in the Appendix. Here are some examples of the 'filtering' process from start to finish in published reviews. I have chosen from those that show a steep decline from initial search to in-depth review.

What is known about successful models of formative assessment for trainee teachers during school experiences and what constitutes effective practice? (Moyles & Yates, 2003):

Initial hits	688
After screening abstracts/titles	233
Keyworded	82
Included in keyword map	58
In-depth review	2

A systematic review of effective literacy teaching in the 4 to 14 age range of mainstream schooling (Hall & Harding, 2003)[18]:

Initial hits	1265
After screening abstract/titles	1187
Keyworded	107
Included in keyword map	80
In-depth review	3

A systematic review of the characteristics of effective foreign language teaching to pupils between the ages 7 and 11 (Driscoll et al, 2004):

Initial hits	5120
After screening abstracts/titles	203
Keyworded	54
In-depth review	4

Supporting pupils with emotional and behavioural difficulties (EBD) in mainstream primary schools: a systematic review of recent research on strategy effectiveness (1999 to 2002) (Harden et al, 2003):

Initial hits	1312
After screening abstracts/titles	233
Keyworded	192
Included in keyword map	55
In-depth review	5

A systematic review of the effectiveness of school-level actions for promoting participation by all students (Dyson et al, 2002):

Initial hits	14692
After screening abstracts/titles	336
Keyworded	41
In-depth review (of 'key studies)	6

A systematic review of the impact of school leadership and management on student outcomes (Bell et al, 2003):

Initial hits	4987
After screening of abstracts/titles (+ additional references obtained)	42
Keyworded & in-depth review	8

Is this sort of thing an embarrassment to the EPPI-Centre? Not really. Oakley touches on the *'relatively low yield of usable studies'* and gives a figure of 0.3 % of initial citations making it through to the in-depth stage in early reviews. The main problem of 'having to search needles to find haystacks', as she sees it, is the cost of time and labour (2003: 27, original emphasis). But the tiny yields of primary studies that are judged methodologically weighty enough for in-depth review serve another, very important purpose within the discourse of systematic review. They allow Oakley and other critics proliferating opportunities to air and extend their 'disappointment' with the quality of educational research. And the authors of these low-yield reviews, faced with the task of justifying conclusions based on a handful of primary studies, and the months of hard work that went into not-reading the rest of the literature, seldom fail to voice their own concerns with the parlous state of their field. Oakley in turn quotes the endorsements of these 'distinguished academics and leaders in their field' (2003: 30).

No time to read…

One might assume that these tiny yields of usable evidence are the unavoidable, if regrettable, result of factors such as narrowly-framed questions or very stringent quality criteria; or indeed of a genuine lack of existing research in the area. And none of these assumptions would necessarily be incorrect. However if you look closely at the instructional literature and at individual reviews, another possible reason emerges – *time.*

Several reviewers allude to difficulties of 'time constraints' or 'manageability' – problems which the EPPI-Centre anticipates. At the data-extraction stage – ie the point at which reviewers would actually begin to read whole texts – the EPPI-Centre Review Companion advises reviewers:

> It is important to bear in mind the time-intensive nature of this stage of the review, since the resources available to a group may determine what it is feasible to undertake in the in-depth review…. If, when you get to this stage of the review you find that you have too many studies to cope with, the EPPI-Centre can help with ways of making the task manageable. (p 15 of 20)

The Companion does not elaborate on what kind of help the Centre offers. But there are glimpses of how some individual review teams handled the problem of making the task manageable.

> Due to the time constraints, electronic databases were used for initial searches. Handsearching was not undertaken (Moyles & Yates, 2003: 13).
>
> Due to the time constraints, electronic databases were used for initial searches and handsearching was not undertaken (Moyles & Stuart, 2003: 13)
>
> Potentially, a full international search was extremely time- and resource-consuming and beyond the capacity of the group to manage. We therefore chose to review literature in English from other countries insofar as it was accessible via standard international databases available in the UK.... It seemed likely that major, funded studies would be recorded in these databases, but that, inevitably, smaller, local studies would not (Dyson et al, 2002: 13).
>
> Only 203 full documents arrived in time; these were screened. The other 75 papers did not arrive in time to be considered in the review (Driscoll et al, 2004: 4).
>
> Due to the time and resource constraints, it was decided that the in-depth review should only include a few studies and that the criteria for the in-depth review should be selected with this in mind (Driscoll et al, 2004: 21)
>
> It was unfortunate that, given the short timescale, we were not able to take account of one group of users, the students themselves (Hall & Harding, 2003: 50).
>
> The extent of user involvement in the review process was less extensive than desired but the time constraints meant that the fast pace of the review process made it difficult to involve users at all stages. (Powell & Todd, 2004: 75)
>
> Due to time constraints, we restricted our choice of journals to those available in the library of our host institution (Harden et al, 2003: 17).[19]

> We did not contact experts in the field due to time constraints (Harden et al, 2003: 18).
>
> Since our review identified a large number of studies, it was not possible to review all of them in depth, so a narrower set of inclusion criteria were developed to select a subset of studies for in-depth review. (Harden et al, 2003: 20)[20]

The application of this 'narrower set of inclusion criteria' left the above reviewers with five studies for in-depth review, which meant that they did not need use their 'reserve' criterion:

> This number was manageable within our timeframe and we did not need to use our 'reserve' criterion which would have enabled us further to exclude studies if they did not evaluate strategies based on a systemic or therapeutic model (Harden et al, 2003: 35)

Slightly less transparently presented in some reviews, then, are a set of additional 'exclusion criteria' that allowed reviewers to keep the corpus of primary studies to a manageable (and in some cases, minute) size. These extra 'exclusion' criteria that have been applied on grounds of time/money could be listed as follows.

- If a journal is not in your institution's library, exclude it.
- If an issue, or even a whole volume is missing, don't try to get it from other sources.
- Don't consult experts.
- Reduce the time spent on the involvement of users.
- Dispense with handsearching.
- Set a cut-off time for getting hold of full copies of studies.
- If too many relevant studies make it through the original exclusion criteria, set some studies aside for a later review.
- Add some additional exclusion criteria.
- And keep an extra one in reserve.

It should be noted that many reviewers did *not* make special appeals to time as an additional constraint on their search or review strategies. Nevertheless, the EPPI-Centre, which exerts strong control over the form and content of reviews, clearly allows such appeals to stand. Oakley, the Director, refers explicitly to the 'less than ideal short-cuts' that may result when the funding for a review runs out (2003: 28). It appears that the Centre does not consider such shortcuts an insurmountable threat to the 'trustworthiness' of the evidence base.

These time-induced strategies for not-reading source material are undoubtedly the product of the Taylorist time-management procedures imposed on reviewers by the systematic review process. Traditional forms of reviewing are subject to time constraints too of course, and these are seldom explicitly acknowledged. But the ruthless prioritising of time and the quantification of effort in the EPPI-framework contributes to a culture

that dismisses intellectual work as mere time-wasting if it cannot conform to schedules and budgets. Any notion of reviewing as involving (amongst other things) slow, careful reading, time for thought, or critique in its wider (non-weighable), sociocultural sense, is stigmatised. As is the idea that reading might actually be pleasurable or personally rewarding. Indeed the whole process of reviewing is often represented as an onerous chore. Here is Oakley, describing the exertions of systematic reviewing.

> Not only is searching for literature a lengthy business, but scrutinising titles and abstracts for relevant information is often exhausting and disappointing, retrieval of hard copies of studies for in-depth review is another time-consuming task, and extracting data from individual studies can take 3 –5 person hours per study. (2003: 28)

Seldom has scholarship been represented in such dispiriting terms. The overwhelming impression from the publications emanating from the EPPI-Centre is that reading is an unreliable, unappetising business that can only be rendered respectable and auditable through hard work and disappointment.

What about *writing?*

Writing is equally degraded in systematic review. As noted above, reviewers are not expected – indeed they appear to be actively discouraged – from any display of rhetorical skill or individual style. 'Synthesis' is represented as a matter of assembling the coded data already entered into the EPPI database and 'interrogating' it using the EPPI software. Writing up must follow 'the EPPI-centre report structure'. This imposes a dreary pattern of identical chapter and section headings, overstuffed methods sections, and overflowing appendices detailing search strategies, coding decisions, application of inclusion criteria, quality weightings etc. Readers have to struggle though pages of empty prose and formulaic phrases that lay out that 'audit trail' that is supposed to lead from conception to conclusion. The ennui induced by these chapters (in this reader at least) contrasts with the introductory 'Background' chapter in most of the reviews, where the authors generally allow themselves to exercise their specialist skills of exposition, argument, conceptualisation and critique according to the generic conventions of traditional academic writing.

Like reading, writing is represented by the EPPI-Centre as not only unhelpfully opaque (which is why reports need to be so closely managed and pre-scripted), but as hard work. The Review Companion, in its final section entitled 'Synthesis and writing the report' rallies flagging reviewers:

> You will be glad to hear that once you get this far, all the hard work is done. The data are present in EPPI-Reviewer, you have a conceptual framework for your synthesis and the software will enable you to interrogate your data in powerful ways (page 17 of 20).

As with reading, the status of writing is diminished to that of an annoying, if minimal chore. The aim, with the aid of the EPPI-Reviewer codes and the software, is to remove the effort and the opacity of writing, and render 'synthesis' as transparent as possible.

Now you see them, now you don't: finding the findings

Perhaps all of this would not matter (though I rather think it still would), if the EPPI-centre methodology was living up to its own claims for transparency, sound evidence, fitness of purpose and value for money. But close reading of the reviews so far produced shows that many of them are rather less ample, informative or cost-effective than proponents of systematic review proclaim. The flaccid bodies of semi-hard evidence that many reviews produce can scarcely merit the considerable costs of producing them, even if you exclude the social cost of demoralising and anathematising a generation of research and researchers. Apart from contributing to that culture of calumny, which some would count a benefit rather than a cost, the other merits of systematic review are small. Few of the education reviews published so far provide evidence that could seriously claim to be trustworthy, based as they are on such small 'samples' of primary studies or such narrowly-defined questions. Even in larger in-depth reviews, the synthesis may include studies that were of low trustworthiness according to the authors' own criteria. [21]

One might expect that this would result in many reviews with little or nothing to report when they arrive at Chapter 5 of the EPPI report structure, 'Findings and Implications'. And some of the conclusions are certainly lukewarm or banal; and many reviews admit the lack of a secure evidential base. However, *in no case does this prevent the reviewers from drawing implications for policy and practice.* For instance, in a review of effective foreign language teaching at ages 7 to 11, the authors state that they are unable to arrive at 'definitive findings'; yet, in the same paragraph, proceed to offer 'the review findings'.

> As there was a dearth of evidence specifically focusing on the characteristics of effective teaching and as the four studies in the in-depth review are concerned with different aspects of teaching, it is not possible to arrive at definitive findings supported by a substantive body of research […] Studies, which did consider teaching, tended to focus on topics such as the qualifications of the teacher or the teaching programme used; this may be useful as background but does not enable conclusions to be drawn about effective teaching. Nonetheless, the review findings are as follows…**(**Driscoll et al, 2004: 4)

A numbered list of 10 findings followed.

Another study, of the impact of ICT on the literacy of learners for whom English is not their first language, describes its limitations thus:

> The primary limitation of the review is that it is small. There are only eight included studies and, of these, none were deemed to provide a 'high' weight of evidence to answer the question at hand; indeed, only three were deemed

> 'medium' […]. The result inevitably remains that it is hard to draw any firm conclusions or draw much in the way of implications for policy (Low & Beverton, 2004: 39).

Yet the authors *do* proceed, on the same page, to present 'more detailed implications', albeit 'with a high degree of caution' (39). In fact they present 4 implications for policy, and 2 for practice.

Even as they lament the size and untrustworthiness of their shrunken bodies of evidence, therefore, many reviewers still seem to feel able (or perhaps obliged) to offer implications for policy and practice. This can lead to uneasy textual manoeuvres and a kind of 'now you see it, now you don't' effect, as the authors try to handle the incommensurable tasks of reporting a lack of evidence while also drawing implications from it. For instance, a review of formative assessment of trainee teachers during school experience contains, on the same page, an acknowledgement of a lack of evidence, and a discussion of what the evidence 'suggests':

> The main limitation of this review is that it cannot confidently recommend, without further research being undertaken, any successful models of formative assessment, nor can it offer any further insight into the debate of what constitutes effective formative assessment practices.
> […]
> Evidence suggests that what appears to be a general move in the UK towards the incorporation of US-style reflective journals within ITT courses is potentially an advantageous move. They appear to increase professional and personal growth and allow teacher educators further insight into the depth of their trainee teachers' pedagogical knowledge (Moyles & Yates, 2003: 38).

To take another example, a review of the impact of headteachers on student outcomes, based on an in-depth review of 8 studies, concludes:

> Effective leadership was confirmed as probably being an important factor in a school's success. (Bell et al, 2003:3)

The tepid nature of this conclusion is doubtless due to the small number of studies on which it was based, and the problematic quality of most of those studies - only one of which seems to have provided wholly trustworthy evidence.[22] Despite this, the authors feel able to commend the results of the review, though the amplitude of their commendation continues to fluctuate:

> [T]he findings broadly confirm the conclusion of the review by Hallinger and Heck (1999) that there is some evidence to support the view that leadership does affect student outcomes, albeit indirectly' (25)

> The overall evidence indicates that headteacher leadership and management does make a difference to pupil performance…' (25)
>
> This review shows that in an area in which significant policy decisions have recently been taken there is very little research evidence available to illuminate the precise nature of the relationship between the leadership and management strategies adopted by headteachers and the learning outcomes of their pupils (26)

The evidence provided by this particular review is hardly 'transparent'. The conclusions seem to derive as much from the expertise and convictions of the reviewers as from the compromised corpus of research on which they are drawing. This does not prevent them, however, from recommending the EPPI-review process as educative:

> Certainly it has shown the need for many to re-assess their ideas about the nature, purposes and scope of literature reviews, the ways in which articles are prepared for publication, and the role of both journal editors and referees. (25)

In common with many of the other reviews published so far, the one 'finding' that the reviewers are able to present unequivocally is the need for the research community to pull up its socks.

It would appear, then, that the embarrassment of ending up with such small yields after so many months of hard labour can result in systematic reviews that jump to conclusions at least as unwarranted as those of so-called 'narrative' reviews. The fact that reviewers are able to reach such conclusions means that the EPPI-Centre process does not eliminate interpretation and judgement from the review process. It just drives them underground. In other words, it renders the review process *less* transparent. One could cite further examples where the 'synthesis' includes rather more than a straightforward adding-up of numerically-weighted evidence. For instance Harlen's discussion of the findings from her in-depth review of 30 primary sources includes reference to fifteen publications that were *not* included in the review (2004: 86-94). This is exactly the kind of interpretive work, based on years of accumulated expertise, that one would expect, and accept, from an experienced academic who is a leading figure in her field. However it does not accord with the strict letter of the EPPI-Centre procedure, since those extra fifteen references have not been properly vetted and weighted for trustworthiness. Similarly, Cordingley et al justify referring to studies assigned a 'low' weight of evidence on the grounds that these were 'used to add texture or as illustration only' (2003: 43). Again, this rationale may be persuasive for some readers, especially in the context of a thoughtful discussion of the different criteria that user groups may bring to the issue of trustworthiness (56-7). But persuasiveness, with its recourse to context, writers' intentions and readers' responses, is not officially included in the EPPI-measuring of quality.

'Synthesis' may involve more, then, than the simple addition or aggregation of evidence culled from the source studies. In fact, I suggest that it must *always* involve more than that. How could it be otherwise? The conversion of weights and numbers back into text – ie into conclusions, findings and implications - is always going to require interpretive and

rhetorical work. Lurking behind the seeming innocence and transparency of data extraction and synthesis are the familiar 'four R's' of scholarly work – reading, writing, reasoning and rhetoric. By denying that these are taking place, systematic review, against its own professed intentions, renders the business of producing evidence less, rather than more open to scrutiny.

Conclusion

Despite its overriding concern with quality, systematic review actually closes down on important questions concerning quality and accountability in educational research. Though its implacable insistence that there is only one right way – the EPPI-way – to address issues of quality, accountability and utility, systematic review makes it more difficult to ask the really difficult questions that many educational researchers would like to ask. By refusing to contemplate alternatives to the audit-trailing of scholarly work, the EPPI-system devalues serious consideration of how quality could be improved without suppressing methodological diversity and critical edge; how researchers could work more meaningfully with users, and how they could respond better to the needs of practitioners and policy-makers, without becoming mere service-providers to these constituencies. Contrary to the assertions of many proponents of evidence-based practice and systematic review, educational researchers often have their own concerns about quality, and are interested in pursuing questions of how to improve it, without sacrificing the diversity of approach, interests, values and purposes on which a democratic research community should be based.

There are some useful things to be learned from systematic review, even as it stands. Reviewers could indeed be more specific about what they looked for, and where. We could all benefit from knowing how to search databases effectively. It might indeed be a good idea for reviewers to give an idea of how convincing they found the evidence (without feeling the need to weigh it). 'Users' certainly can productively contribute to the framing of review questions and to other stages of the review process, and there are interesting accounts in the reviews of the range of forms that user involvement can take. Like other national policies – for example, the National Literacy Strategy – there is stuff in there that many professionals and researchers would consider good practice. This is one reason why dissent is so easily framed as dishonourable. ('Oh, so you *don't* think all children are entitled to learn to read?' 'So you *don't* think that reviewers should be answerable for their methods?').

But the overarching framework that bends research to the wheel of accountability counteracts the good practice, especially since systematic review is only one component of a much larger piece of policy machinery that is rolling in one direction. Think again of the National Literacy Strategy. Like systematic review, the NLS also involves prescriptive vocabularies ('bookbanking', 'scribing', 'powerful verbs' 'strategy-checks' etc), broken-down practice (each literacy hour broken down into segments; each task broken down into pre-set objectives), and ruthless regimentation and monitoring (only one right way to write a recipe or read a poem). Comparison could also be made with Ofsted school inspection, where the regulation of reporting similarly imposes conformity to a ponderous structure of categories, section headings and gradings, and a bland house

style that permits inspectors, like reviewers, only to report what falls within the category system, in a voice that is evacuated of interest or equivocation. Systematic review is just one part, I suggest, of a pervasive attempt to reconfigure and regulate professional and academic practices and identities by acting on the very words that people speak and write. It seems to be based on assumption that discipline, in its Foucauldian sense, can be made to work 'backwards': from word to thought, from speech to self.

The intellectual, social and political implications of this are malign. If the project of disabling critique and disciplining academic work succeeds, the outcome is likely to be a diminution in the social usefulness of research knowledge, the continuing oversimplification of educational problems and solutions, and a less well-educated, less critical community of researchers. It might be tempting to dismiss the EPPI-phenomenon as a small, cultish alternative to a mainstream academic culture that can easily shrug off its impertinences, or bring it round to a more informed understanding of what can and cannot be claimed by an evidence-based approach. And 'behind' some of the reports one can indeed detect traces of attempts to mitigate the extremity of systematic review. Some authors do discuss the difficulties and potentially adverse effects of the approach.[23] It is entirely possible that some reviewers successfully fight their corner with EPPI-Centre staff over issues of principle or procedure, or manage to smuggle some nuance or complexity into their reports. Moreover, as I suggested above, behind the cover story of transparent, interpretation-free evidence accumulation, the traditional work of interpretation and argument goes on. But these are all covert operations, carried out under cover of endorsement of the system in principle; and they rely on an existing body of experienced researchers who know that there are other ways to do things. [24] As long as the EPPI-Centre and its supporters continue to promote the official story of reading-lite reviewing and interpretation-free evidence, and to sell it to novice researchers, a dangerously over-simplified version of research and scholarship is continuously recycled.

Already systematic review is making inroads into new academic territories. It is now on sale as a taught course or a 20-credit module towards a masters degree at the University of London, and is therefore convertible into the currency of academic qualifications from a high-status university.[25] It is increasingly becoming a requirement in government contracts, and a yard-stick by which to measure the quality of proposed research designs (cf Torrance, 2004). The Funders Forum of the National Educational Research Forum (NERF) identifies a 'systematic review skills deficit' and advises commissioners to employ only those properly trained in systematic review procedures.[26] Finally, systematic review is snapping at the heels of journal editors. Supporters of the EPPI system are convening meetings at which editors are exhorted to change their abstracting guidelines to facilitate electronic searching and help prevent unnecessary reading (NERF, 2004). Peer review is under attack, and referees and editors who reject papers based on systematic review have been challenged on the grounds that, if it's good enough for the DfES and the EPPI-Centre, it's good enough for a research journal (Stronach, 2004).

Like all accountability systems, systematic review is hostile to anything that cannot be seen, and therefore controlled, counted and quality assured. Reading and writing are deeply problematic from that angle, because things happen out of sight, and outcomes are

never wholly predictable. Different readers will extract differing meanings from texts, according to their background assumptions and their current priorities and beliefs. Understanding and interpretation leave no traces while they are happening. The only possible access one can have to them is through texts, which are, of course, themselves subject to further acts of interpretation. Reading and writing *always* run the risk of opening up, or calling attention to, the calamitous gaps that exist in the foundations of thought, knowledge and principled action. Writing inevitably opens the gap between representation and reality. Reading does the same for facts and interpretation. As Derrida has shown (eg 1980, 1981), this is why they have posed a threat to so many disciplines and cultural practices. This is why the rage for transparency – for unclouded access to the fundamentals that are concealed as well as carried by texts – is a very old, as well as a very contemporary phenomenon (see MacLure, 2003: Ch 6). Much contemporary education policy seems to be animated by this fear and distrust: not just of the unreliability of professionals, but of *language itself*. Policy seems to be suffering an acute case of that old ontological panic about 'signs' and the unreliable access that they offer to the stuff that is supposed to lie behind them – truth, knowledge, meaning, evidence, standards.

The EPPI-Centre's response to the threat of reading/writing is not so much a rage, though, as a childish tantrum. It amounts to a continuous, foot-stamping whine about the impasse between texts and truths, and the indirection, ambiguity, contestability, unpredictability and uncertainty that are the price we pay for what we know. In waging war on evidence that can never fully be prised out of its textual embodiment, and on academics whose work resists counting and quality-control, the EPPI-Centre approach is brutishly destructive of some of the most important aspects of research and scholarship – ie the reading, writing, thinking, interpreting, arguing and justifying - out of which knowledge is precariously produced.

REFERENCES

Systematic reviews referred to in the text

All reviews are published in: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, accessible online at: http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?&page=/reel/reviews.htm

Andrews R, Burn A, Leach J, Locke T, Low G, Torgerson C (2002). A systematic review of the impact of networked ICT on 5-16 year olds' literacy in English (EPPI-Centre Review, version 1.1).

Andrews R, Torgerson C, Beverton S, Locke T, Low G, Robinson A, Zhu D (2004) The effect of grammar teaching (syntax) in English on 5 to 16 year olds' accuracy and quality in written composition.

Bell L, Bolam R, Cubillo L (2003) A systematic review of the impact of school leadership and management on student outcomes.

Bennett J, Hogarth S, Lubben F (2003) A systematic review of the effects of context-based and Science-Technology-Society (STS) approaches in the teaching of secondary science. Version 1.1.

Bennett J, Lubben F, Hogarth S, Campbell B (2004) A systematic review of the use of small-group discussions in science teaching with students aged 11-18, and their effects on students' understanding in science or attitude to science.

Burn A, Leach J (2004) A systematic review of the impact of ICT on the learning of literacies associated with moving image texts in English, 5-16.

Cordingley P, Bell M, Rundell B, Evans D (2003) The impact of collaborative CPD on classroom teaching and learning.

Deakin Crick R, Coates M, Taylor M, Ritchie S (2004) A systematic review of the impact of citizenship education on the provision of schooling.

Driscoll P, Jones J, Martin C, Graham-Matheson L, Dismore H, Sykes R (2004) A systematic review of the characteristics of effective foreign language teaching to pupils between the ages 7 and 11.

Dyson A, Howes A, Roberts B, (2002). A systematic review of the effectiveness of school-level actions for promoting participation by all students (EPPI-Centre Review, version 1.1).

EPPI Centre (2001) (Gough, D, Oliver, S, Brunton, G, Selai, C & Schaumberg, H) for the

Department of the Environment, Transport and the Regions). The effect of travel modes on children's mental health, cognitive and social development; a systematic review.

Evans J, Harden A, Thomas J, Benefield P (2003) Support for pupils with emotional and behavioural difficulties (EBD) in mainstream primary classrooms: a systematic review of the effectiveness of interventions.

Fletcher M, Lockhart I (2003) The impact of financial circumstances on engagement with post-16 learning: a systematic map of research (EPPI-Centre Review, version 1).

Francis B, Skelton C, Archer L (2002). A systematic review of classroom strategies for reducing stereotypical gender constructions among girls and boys in mixed–sex UK primary schools (EPPI-Centre Review, version 1.1).

Gough DA, Kiwan D, Sutcliffe K, Simpson D, Houghton N (2003). A systematic map and synthesis review of the effectiveness of personal development planning for improving student learning.

Hall K, Harding A (2003) A systematic review of effective literacy teaching in the 4 to 14 age range of mainstream schooling.

Harden A, Thomas J, Evans J, Scanlon M, Sinclair J (2003) Supporting pupils with emotional and behavioural difficulties (EBD) in mainstream primary schools: a systematic review of recent research on strategy effectiveness (1999 to 2002).

Harlen W (2004) A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes.

Harlen W, Deakin Crick R (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1).

Harlen W, Deakin Crick R (2003) A systematic review of the impact on students and teachers of the use of ICT for assessment of creative and critical thinking skills.

Howes A, Farrell P, Kaplan I, Moss S (2003) The
impact of paid adult support on the participation and learning of pupils in mainstream schools.

Locke T, Andrews R (2004) A systematic review of the impact of ICT on literature-related literacies in English 5-16.

Low G, Beverton S (2004) A systematic review of the impact of ICT on literacy learning in English of learners between 5 and 16, for whom English is a second or additional language.

Moyles J, Stuart D (2003) Which school-based elements of partnership in initial teacher training in the UK support trainee teachers'professional development?

Moyles J, Yates R (2003) What is known about successful models of formative assessment for trainee teachers during school experiences and what constitutes effective practice?

Parker-Jenkins M, Hewitt D, Brownhill S, Sanders T (2004) What strategies can be used by initial teacher training providers, trainees and newly qualified teachers to raise the attainment of pupils from culturally diverse backgrounds?

Penn H, Barreau S, Butterworth L, Lloyd E, Moyles J, Potter S, Sayeed R (2004) What is the impact of out-of-home integrated care and education settings on children aged 0-6 and their parents?

Powell S, Tod J (2004) A systematic review of how theories explain learning behaviour in school contexts.

Scheurich, J. J (2000) A RoUGH, ramBling, strAnGe, muDDy, CONfusing, e1LIPtical Kut: From an archaeology of plain talk, *Qualitative Inquiry*, 6, 3: 337-348.

Torgerson C, Zhu D (2003) A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English, 5-16.

Totterdell M, Woodroffe L, Bubb S, Hanrahan K (2004) The impact of NQT induction programmes on the enhancement of teacher expertise, professional development, job satisfaction or retention rates: a systematic review of research on induction.

References to other work

Avis, J. (2003) Work-based knowledge, evidence-informed practice and education, *British Journal of Educational Studies,* 51, 4: 369-389.

Breton, A. (1969) *Manifestoes of Surrealism.* (Trans R. Seaver & H R Lane). Michigan: University of Michigan Press.

Derrida, J. (1980) *The Archaeology of the Frivolous. Reading Condillac.* Pittsburgh: Duquesne University Press.

Derrida, J. (1981) *Dissemination,* trans, with an Introduction by B. Johnson. London: Athlone Press.

Derrida, J. (1988) *Limited Inc.* Evanston, Chicago: North Western.

EPPI-Centre (2003) EPPI-Centre Review Companion. http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?page=/reel/tools.htm

EPPI-Centre (nd) What is a Systematic Review? http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?page=/reel/about_reviews.htm

Evans, J. & Benenfield, P. (2001) Systematic reviews of educational research: does the medical model fit? *British Educational Research Journal,* 27, 5: 527-541.

Elliott, J. (2001) Making evidence-based practice educational, *British Educational Research Journal,* 27, 5: 555-574.

Fairclough, N. (1989). *Language and power*. London: Longman.

Foucault, M. (1980) *Power/Knowledge: Selected Interviews and Other Writings 1972-77 by Michel Foucault,* ed. C. Gordon. London: Harvester Wheatsheaf.

Giroux, H. (1992) Language, difference and curriculum theory: beyond the politics of clarity, *Theory into Practice,* 31, 219-227.

Greenblatt, S. (1991) *Marvellous Possessions: The Wonder of the New World.* Oxford: Clarendon.

Hammersley, M. (2001) On 'systematic' reviews of research literatures: a 'narrative' response to Evans and Benenfield, *British Educational Research Journal,* 27, 5: 543-554.

Hargreaves, D. (1996) Teaching as a research-based profession: possibilities and prospects. Teacher Training Agency Annual Lecture. London: TTA.

Hillage, J, Pearson, R, Anderson, A & Tamkin, P. (1998) *Excellence in Research in Schools.* London: Department for Education & Employment/Institute of Employment Studies.

Kristeva, J. (1982) *Powers of Horror: An Essay on Abjection.* New York: Columbia University Press.

MacLure, M. (2003) *Discourse in Educational and Social Research.* Milton Keynes: Open University Press.

NERF (National Educational Research Forum) (2004) Quality of research reporting and abstracts: Judy Sebba's presentation, in Report from the Third Education Journal Editors' Seminar. NERF: http://www.nerf-uk.org/themes/keyuse/intermediaries/publishers/?version=1

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

Strathern, M. (2000) The tyranny of transparency, *British Educational Research Journal,* 26, 3: 309-201.

Stronach, I. (2004) Deconstructing 'quality' and reconstructing critique. Paper presented to the Annual Conference of the British Educational Research Association, Manchester, September.

Thomas, G. & Pring, R. (eds) (2003) *Evidence-Based Practice in Education.* Buckingham: Open University Press.

Tooley, J & Darby, D. (1998) *Educational Research: A Critique.* London: Office for Standards in Education.

Torrance, H. (2004) Systematic reviewing – the 'call centre' version of research synthesis. Time for a more flexible approach. Paper presented to the ESRC/RCBN Seminar on Systematic Reviewing, University of Sheffield, 24 June 2004.

Zeller, N. and Farmer, F.M. (1999) 'Catchy, clever titles are not acceptable': style, APA, and qualitative reporting, *International Journal of Qualitative Studies in Education,* 12 (1): 3-20

Appendix.

Number of primary studies subjected to 'in-depth review' in 28 systematic reviews published on the EPPI-Centre 'REEL' website in September 2004.

Review	No. of studies reviewed 'in-depth'
Moyles & Stuart (2003)	2
Moyles & Yates (2003)	2
Totterdell et al (2004)	2
Hall & Harding (2003)	3
Driscoll et al (2004)	4
Powell (2004)	5
Parker-Jenkins et al (2004)	5
Bennett et al (2003)	5
Harden et al (2003)	5
Dyson et al (2002)	6
Locke & Andrews (2004)	7
Low & Beverton (2004)	8
Bell et al (2003)	8
Burn & Leach 2004)	9
Francis et al (2002)	9
Penn et al (2004)	9
Andrews et al (2004)	10
Harlen & Deakin Crick (2003)	12
Torgerson & Zhu (2003)	12

Notes

[1] References for opening quotations: Breton (1969: 6); Montaigne (quoted in Greenblatt, 1991: 146-7)

[2]The full title of the EPPI-Centre is The Evidence for Policy and Practice Information and Co-ordinating Centre (Director: Ann Oakley). The Centre's homepage is http://eppi.ioe.ac.uk/EPPIWeb/home.aspx

[3] This is a version of a paper of the same title presented to the Annual Conference of the British Educational Research Association, Manchester, September, 2004. The paper was part of a symposium entitled 'Quality Street and other cul-de-sacs: what is 'quality' in education and educational research, and where are its cutting edges'?

[4] See for example Avis (2003); Elliott (2001); Torrance (2004) and chapters in the second half of Thomas & Pring (2003).

[5] Cf Strathern (2000); Avis (2003); Hammersley (2001).

[6] 'What is a Systematic Review', http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?page=/reel/about_reviews.htm

[7] Cf. Oakley: 'the disadvantages of not doing this [ie introducing more 'codification into the knowledge base'] can … literally be fatal' (2003: 22). Oakley is referring here to 'areas such as HIV/AIDS education' but the remark comes in the middle of a paragraph about the general importance of making it more difficult for health and education professionals to 'hide failures'.

[8] The original critiques are: Hargreaves (1996), Hillage et al (1998) and Tooley & Darby (1998).

[9] 'How will the [review] group ensure that the methodological and conceptual issues are understood and agreed upon across the review team in time for the protocol to be written?' 'Which computer software are you going to use to manage citations?' 'What will happen if this person is unable to fulfil this role'? The 'Quality Assurance' questions in the Companion are written in red typeface and are presumably especially important: 'Please describe the procedures by which the EPPI-Centre worked with the group to provide external quality assurance of application of inclusion/exclusion criteria'. 'Please describe the procedures by which the EPPI-Centre worked with the group to provide external quality assurance of application of mapping keywords'. 'Please describe the procedures by which the EPPI-Centre worked with the group to provide external quality assurance of data extraction'. 'EPPI-Centre Review Companion', *passim*: http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?page=/reel/tools.htm

[10] One review team considered (but decided against) departing from the EPPI-Centre's framework (Moyles & Yates, 2003: 39). Another was somewhat equivocal, describing systematic review approvingly as 'a searchlight', but also debating its inflexibility and narrow focus, and its inability to capture the conceptual complexity of their field (Dyson et al, 2002: 53, 54-55). There were also instances of reviewers using their specialist knowledge and skills in ways that went beyond the remit of mere 'synthesis', to incorporate old-school, 'narrative' reviewing habits such as interpretation and argument: see further below.

[11] There are also strange formations: eg to 'data extract' ('Two members of the EPPI-Centre also data-extracted a sample of papers as part of the quality assurance process' (Cordingley et al, 2003:25) or 'mapping keywording' ('Who will be mapping keywording the studies?' asks the EPPI-Centre Review Companion).

[12] See for example Derrida (1981), Zeller & Farmer (1999). MacLure (2003, Ch 6) provides an overview of the 'threat of writing', or textuality, and the ways in which philosophers, literary theorists and social scientists have attempted to contain that threat. Systematic review is a particularly extreme reaction to the threat of textuality: see 'Conclusion'.

[13] 'How much time / how many studies can [the named reviewer] spend / do?' (EPPI-Centre Review Companion: 16/20)

[14] Not all reviews start out with narrow questions. For instance Dyson et al (2002) and Cordingley et al (2003) both began with broad questions, and their initial searches yielded very large numbers of potentially relevant 'hits'. However this initial breadth had narrowed dramatically by the final, 'in-depth review' stage: from 14,692 to 6, and 13,479 to 17 respectively: see below.

[15] Reviewers will, of course, have read those research studies with which they were already familiar, unless the review team are novices to the substantive area – a possibility that is not ruled out in systematic review (cf Torrance, 2004).

[16] Some reviews carried out their in-depth review on the total corpus of studies included in the keyword map: Harlen (2004: 30 studies); Harlen & Deakin Crick (2003: 12 studies); Deakin Crick et al (2004: 14 studies); Howes et al (2003: 24 studies)

[17] A total of 30 education reviews were published on the EPPI-Centre website at this date. However two of these did not report an 'in-depth' review stage: EPPI-Centre (2001) and Fletcher & Lockhart (2003).The format of these two reviews also diverged in other respects from the usual EPPI review structure.

[18] 3 primary studies were 'subjected to the full EPPI procedures of in-depth reviewing' (Hall & Harding, 2003: 30). There is some ambiguity: the reviewers identified a total of 12 studies for in-depth review on the grounds of their 'direct relevance to the Teacher Training Agency'; but only three were subjected to the full treatment because only these gained high enough ratings for quality of empirical evidence.

[19] The resulting exclusions included one presumably key journal, since it was listed by the authors in their protocol (J Emotional & Behavioural Disorders). Moreover if issues, or even whole volumes, were missing from their institution's library, the authors did not try to obtain these from other sources. 14 issues were 'missing', and 17.5 whole volumes. (Harden et al, 2003: 72)

[20] The 'narrower set' of inclusion criteria applied by Harden et al (2003) excluded all studies that did not employ a control or comparison group or a reversal design, or which employed a sample size of less than 20 - thus endorsing the bias towards quantitative design as the benchmark of quality in systematic review.

[21] For example, Evans et al report an in-depth review of 27 primary studies (placing them near the top of the 'league table': see Appendix). However, only 10 of these were judged to be 'methodologically "sound"' (2003: 4). Cordingley et al rated only 2 studies as 'high' on one of their two sub-questions, and *none at all* on the other (2003: 43). Low & Beverton also rated none of their 8 studies as 'high', and only 3 as 'medium' (2004: 39).

[22] Of the eight studies (which are not assigned numerical 'weights' according to the usual EPPI-procedure), one was 'of limited relevance' (Bell et al, 2003: 17); one was a case study of a single school, from which 'one cannot generalise' (22); one showed 'moderate' and 'negligible' correlations between leadership and student outcomes (17); one found no 'significant, positive relationship' (18); one found that principals made 'a disappointing contribution to student engagement' (19); one found 'indirect' effects (19); one had a faulty design and, in any case, found only a 'weak correlation between leadership and achievement' (20); and one found a 'mediated' (ie indirect) relationship between 'transformational leadership' and student achievement in Maths (20).

[23]For instance Dyson et al discuss the features of systematic review that tend towards narrowing of focus and reduced flexibility, referring to the '(understandable) emphasis on tight delineation of review topic, requirement for clear a priori criteria for what forms of evidence can be included and current lack of procedures for 'analogical' synthesis.' They also note 'the practical issue of managing a very wide ranging review within a limited timescale and budget' (2002: 55).

[24]Amanda Coffey, as Discussant to the symposium in which the original version of this paper was presented, cautioned against manufacturing a 'moral panic' about systematic review. I take her point. It may be that this paper over-estimates the power of the evidence 'movement' and its ability to influence research and scholarship. I think I would still argue that the time has come for concerted opposition rather than covert operations. But I could be wrong.

[25] Details of the course can be found online by following the link on the EPPI-Centre's homepage at http://eppi.ioe.ac.uk/EPPIWeb/home.aspx

[26] http://www.nerf-uk.org/funders/systematic/