

THE MAGENTA BOOK

Chapter 2: What do we already know

Background Document

Dr Phil Davies

Government Chief Social Researcher's Office
Cabinet Office Strategy Unit

June 2003

1.0 The Problem

1.1 An essential first step in planning a policy evaluation is to determine what is already known about the topic in question from the full range of existing evidence.

This is important for at least four reasons:

- i. It may be that there is already sufficient evidence on the likely effectiveness of a policy, programme or project so that further primary evaluation is unnecessary. Such a situation is very unlikely for the reasons outlined below.
- ii. The existing evidence may be ambiguous, inconclusive, or of uncertain quality indicating that further evaluation is necessary and that specific aspects of the policy in question need addressing.
- iii. It may be that there is no valid, reliable and relevant evidence available at all on the policy in question. This will help determine the nature and scope of the evaluation that needs to be undertaken.
- iv. Any single evaluative study may illuminate only one part of a policy issue, or its findings may be sample specific, time specific, or context specific. This makes it difficult to establish the generalisability and transferability of findings from existing research evidence which, in turn, will influence what requires evaluating.

1.2 Establishing what is already known about a policy, programme or project, however, presents a major challenge for knowledge management. The sheer amount of potential research evidence in most substantive areas of social science and public policy, coupled with the rapid growth of access to knowledge and information as a result of information technology, make it almost impossible to

keep abreast of the research literature in any one area. Given the limitations of humans' information processing abilities, the complexity of modern professional life almost certainly exceeds the capacity of the *unaided* human mind (Eddy 1999).

1.3 The problems of information overload are compounded by the fact that not all research and information is of equal value. Variations in the quality of primary studies, reporting practices, standards of journal indexing and editing, and publication criteria mean that the existing research literature is often of variable quality. Consequently, seemingly similar studies may be of different focus, value and relevance to users of research evidence. Some way of differentiating between high and lower quality studies, as well as relevant and irrelevant evidence, is required.

2.0 A Solution

2.1 Systematic reviews of existing research literature are increasingly being used as a valid and reliable means of harnessing the existing research evidence. They can also allow a cumulative view of existing research evidence to be established. As Cooper and Hedges (1994:4) point out, systematic reviews “attempt to discover the consistencies and account for the variability in similar-appearing studies”. Also, “seeking generalisations also involves seeking the limits and modifiers of generalisations” (*ibid*) and, thereby, identifying the contextual-specificity of available research evidence.

3.0 How Is This Different From What Is Normally Done?

3.1 Systematic reviews differ from other types of research synthesis (e.g. narrative reviews and vote counting reviews) by:

- being more systematic and rigorous in the ways they search and find existing evidence;
- having explicit and transparent criteria for appraising the quality of existing research evidence, especially identifying and controlling for different types of bias in existing studies;
- having explicit ways of establishing the comparability (or incomparability) of different studies and, thereby, of combining and establishing a cumulative view of what the existing evidence is telling us.

3.2 Two common methods of synthesising existing evidence are *narrative reviews* and *vote counting reviews*.

3.3 *Narrative Reviews*

The simplest form of research synthesis is the traditional qualitative literature review, often referred to as the *narrative review*. Narrative reviews typically attempt to identify:

- readily available literature on a subject or topic;
- which methodologies have been used in that literature;
- what samples or populations have been studied (and not studied);
- what findings have been established;
- what caveats, qualifications and limitations exist in the available literature.

3.4 Narrative reviews may (or may not) provide an overview or summary of research on a topic. More typically they identify the range and diversity of the available literature, much of which will be inconsistent or inconclusive.

3.5 A major limitation of narrative reviews is that they are almost always *selective*. They do not always involve a *systematic, rigorous and exhaustive* search of *all* the relevant literature using electronic and print sources as well as hand searching and ways of identifying the ‘grey’ literature (i.e unpublished studies or work in progress). This means that traditional narrative literature reviews often involve *selection bias* and/or *publication bias*. The latter is a consequence of some journals disproportionately reporting studies with positive outcomes, whilst some other sources disproportionately report studies with negative outcomes.

3.6 Narrative literature reviews are also often *opportunistic* in that they review only literature and evidence that is readily available to the researcher (the file drawer phenomenon). Some narrative reviews may discard studies that use methodologies in which the researcher has little or no interest. Alternatively, they may include studies that use different methodologies and which do not lend themselves to meaningful comparison or aggregation. Narrative reviews often provide few details of the procedures by which the reviewed literature has been identified and appraised. It is also often unclear how the conclusions of narrative reviews follow from the evidence presented. This lack of transparency makes it difficult to determine the selection bias and publication bias of narrative reviews,

and runs the risk of over-estimating (or in some cases under-estimating) the effectiveness of interventions in ways that are hard to identify.

3.7 With systematic reviews the problems of selection bias and publication bias are dealt with by identifying and critically appraising *all* of the available research literature, published and unpublished. This involves detailed hand searching of journals, textbooks, and conference proceedings, as well as exhaustive electronic searching of the existing research literature.

3.8 Systematic reviews also differ from narrative reviews in that they make explicit the search procedures for identifying the available literature, and the procedures by which this literature is critically appraised and interpreted. This affords a degree of transparency by which other researchers, readers and users of systematic reviews can determine what evidence has been reviewed, how it has been critically appraised, and how it has been interpreted and presented. This, in turn, allows for other interpretations of the evidence to be generated, and for additional studies of comparable quality to be added to the review, if and when they become available. In these ways, an interactive and cumulative body of sound evidence can be developed.

3.9 *Vote Counting Reviews*

A type of research synthesis that attempts to be cumulative is the *vote counting review*. This attempts to accumulate the results of a collection of relevant studies by counting “how many results are statistically significant in one direction, how many are neutral (i.e. “no effect”), and how many are statistically significant in the other direction” (Cook *et al*, 1992:4). The category that has the most counts, or votes, is taken to represent the modal or typical finding, thereby indicating the most effective means of intervention.

3.10 An obvious problem with voting counting reviews is that they do not take into account the fact that some studies are methodologically superior than others and, consequently, deserve special weighting. Systematic reviews differentiate between studies of greater and lesser sample size, power and precision and weight them accordingly. [To see how such weighting of different studies is done see Deeks, Altman and Bradburn, (2001)].

3.11 Another problem with vote counting reviews is that they fail to indicate “the possibility that a treatment might have different consequences under different conditions” (Cook *et al*, 1992:4). Crude counting of studies in terms of the direction of outcomes does not take into account that “person and setting factors are especially likely to moderate causal relationships and help explain why a treatment has the effects it does” (Cook *et al*, 1992:22). Systematic reviews attempt to incorporate such contextual factors by closely analysing the findings and limitations of different studies and identifying their implications for policy and

practice. Where there is evidence on a topic from qualitative research this can also be used to identify important contextual and mediating factors.

4.0 What is Meta-Analysis?

4.1 Meta-analysis is a type of systematic review that aggregates the findings of comparable studies and “combines the individual study treatment effects into a “pooled” treatment effect for all studies combined” (Morton, 1999). The term ‘meta-analysis’ has been commonly attributed to Gene Glass (1976) who used the term to refer to “the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings”. The statistical basis of meta-analysis, however, can be traced back to seventeenth century astronomy, which suggested “that combinations of data might be better than attempts to choose amongst them” (Egger, Davey-Smith and O’Rourke, 2001:8)

4.2 In the two decades or more since Glass’s original meta-analytic work on psychotherapy (Smith, Glass and Miller, 1980) and class size (Glass and Smith, 1979; Smith and Glass, 1980; Glass, Cahen, Smith and Filby, 1982), meta-analysis has developed considerably in terms of the range and sophistication of data-pooling and statistical analysis of independent studies (see Kulik and Kulik, 1989, Cook *et al*, 1992, Cooper and Hedges, 1994 and Egger, Davey Smith, and Altman, 2001 for more detailed accounts of these developments). They have also been undertaken in substantive areas other than education and health care, including criminology, social work and social welfare.

4.3 Meta-analysis is perhaps best known for combining the results of randomised controlled trials (see Chapter 6), though as Egger, Davey-Smith and Schneider (2001:211) point out they are also commonly undertaken on non-randomised data from primary studies that use case-control, cross-sectional, and cohort designs. Non-randomised studies, however, are much more susceptible to the influence of confounding factors and bias and may “provide spuriously precise, but biased, estimates of association” (*ibid*).

4.4 Meta-analysis of randomised controlled trials, on the other hand, assumes that each individual trial provides an unbiased estimate of the effects of an experimental intervention, and that any variability of results between studies can be attributed to random variation. Consequently, by combining the results of randomised controlled trials an overall effect of the intervention can be estimated that is unbiased and has measurable precision.

5.0 What Are The Limitations of Meta-Analysis?

5.1 The “Apples and Pears” Problem

Meta-analysis has its own limitations. Like other types of research synthesis it requires *focussed* questions to be asked about:

- the intervention(s) under investigation
- the population (or sub-groups) studied
- the outcomes that are being assessed.

Given that each of these factors may vary across individual studies this presents a challenge for the meta-analyst to ensure that there is real consistency between primary studies on all three dimensions. If this is not done there is the “apples and pears” problem of falsely aggregating studies that are not really comparable.

5.2 There are ways of testing whether or not different primary studies are sufficiently similar (or homogeneous) for their findings to be aggregated into a pooled estimate of overall effect size. Funnel plot analysis is one such method of testing for homogeneity or heterogeneity of different primary studies (see Deeks, Altman and Bradburn, 2001). Moreover, there are different ways of analysing studies where there is greater homogeneity (i.e. using fixed effects models) and where there is greater heterogeneity (i.e. using random effects models). For further discussion of random-effects and fixed-effects models of meta-analysis see Hedges (1994), Raudenbusch (1994) and Deeks, Altman and Bradburn, (2001).

5.3 *The Adequacy of Searching*

Meta-analysis may also be limited by the degree of systematic and comprehensive searching that is undertaken for relevant and appropriate primary studies. Extensive, if not exhaustive, searches are required of databases, textbooks, journals, conference proceedings, dissertation abstracts, and research-in-progress, using electronic and hand searching methods. The need to search unpublished sources (including research-in-progress) is crucial given the problems of positive (and in some cases negative) publication bias in journals and other print sources.

5.4 *The Quality of Primary Studies*

Meta-analysis requires high quality standards of methodology and reporting in the primary studies being synthesised. These include:

- The validity and reliability of tests and outcome measures used in the primary studies;
- Follow-up data that are consistent with baseline data;
- The reporting of (and accounting for) participants lost-to-follow-up at different data collection points (i.e. attrition bias);
- Accounting for missing data on moderator and mediating variables;
- Systematic differences in the treatment of comparison groups other than the intervention under investigation (i.e. performance bias)
- The appropriateness of descriptive and inferential statistics (means and standard deviations, chi-squares, odds ratio and confidence intervals) used in the primary studies;
- The types of statistical manipulation (e.g. logarithmic transformations of data) used in the primary studies;
- Ensuring the independence of primary studies so that the results of individual studies are not included more than once in a meta-analysis, thereby double counting studies in estimating the effect size.

5.5 As Cook *et al* (1992) point out, there are several ways in which problems of inadequate statistical reporting can be handled by meta-analysts. These include:

- using external sources to establish the validity and reliability of instruments used in primary studies;
- contacting the primary investigator(s) to obtain additional data or clarification of procedures used;
- reporting deficiencies of primary data in the meta-analysis, thereby distinguishing between good and poor data.

5.6 All of these problems need to be confronted and resolved by meta-analysts in order to provide unbiased estimates of the overall likely effects of an intervention and greater precision than that given by narrative or vote counting reviews.

6.0 What About the Synthesis of Non-Experimental Studies?

6.1 Methods for synthesising data from primary studies that use experimental (randomised control trials) and quasi-experimental methods (such as case-control, cross-sectional, and cohort designs) are well developed. The synthesis of primary studies that use non-experimental methods such as in-depth interviews, observational methods, participant observation and ethnography is currently less developed. Work is currently underway to determine quality criteria for these methods of qualitative research and evaluation, and to develop a framework for the critical appraisal of qualitative evaluation studies (Cabinet Office, 2002). Procedures for undertaking systematic reviews of different types of evidence have been developed (EPPI-Centre, 2001), and methods for synthesising qualitative

research are being developed (Oakley, Gough and Harden, personal communication).

6.2 Two earlier attempts to develop the synthesis of non-experimental studies include *meta-ethnography* and *best evidence synthesis*.

6.3 *Meta-Ethnography*

Meta-ethnography attempts to summarise and synthesise the findings of qualitative studies, especially ethnographies and interpretive studies. Meta-ethnography claims to “be interpretive rather than aggregative” (Noblit and Hare, 1988:11). and covers:

research that is termed ethnographic, interactive, qualitative, naturalistic, hermeneutic, or phenomenological. All these types of research are interpretive in that they seek an explanation for social or cultural events based upon the perspectives and experiences of the people being studied. In this way, all interpretive research is “grounded” in the everyday lives of people.
(Noblit and Hare, 1988:12)

6.4 Like meta-analysis, meta-ethnography “seeks to go beyond single accounts” (Noblit and Hare, 1988:13), but instead of doing so by aggregating samples and identifying consistencies and variability between different studies, it does this by “constructing interpretations, not analyses” and by revealing “the analogies between the accounts” (*ibid*).

6.5 Meta-ethnography “reduces the accounts while preserving the sense of the account through the selection of key metaphors and organisers” (*ibid*). These refer to “what others may call themes, perspectives, organisers, and/or concepts

revealed by qualitative studies” (*op cit*: 14). To this extent, meta-ethnography would appear to have more in common with narrative reviews than with vote counting reviews and meta-analyses.

6.6 *What are the Problems and Limitations of Meta-Ethnography?*

Meta-ethnography has some of the same problems as meta-analysis and other types of research synthesis. These include:

- establishing criteria for which studies to include and exclude in a meta-ethnographic review.
- handling the diversity of the questions being asked and the theoretical perspectives from which these questions are generated;
- dealing with the heterogeneity of primary studies that use qualitative research and evaluation methods;
- balancing summary statements of qualitative studies with their contextual specificity;
- providing overviews of qualitative studies without ignoring the “meaning in context” and the “ethnographic uniqueness” that is central to ethnographic and qualitative inquiry.

6.7 Meta-ethnography is seen by those who require quantitative synthesis of existing evidence as being limited by its inability:

- to provide statistical accumulation of findings;

- to allow prediction or to specify any degree of confidence about qualitative findings;
- to allow for the statistical control of bias.
- to test for, and control, the heterogeneity/homogeneity of different studies,

6.8 These latter concerns about the synthesis of qualitative studies, however, seem to miss the point of what ethnographies and other qualitative studies are trying to achieve (Davies, 2000). That is, to provide rich descriptions of naturally occurring activity, rather than experimental constructions, and to interpret the individual and shared meanings of the topic under investigation for different individuals and groups. Qualitative inquiry is particularly concerned with the contextual specificity of these meanings rather than with their de-contextualised generalisability. To subject such findings to statistical representation, manipulation and testing is usually inappropriate, other than to identify patterns of consistency and inconsistency amongst different qualitative studies.

6.9 *Best Evidence Synthesis*

Slavin (1984, 1986) has proposed that the type of methods used to generate research evidence is less important than the quality of the primary studies undertaken, whatever methodological approaches are used. Slavin suggests that what is required is ‘best evidence synthesis’ in which “reviewers apply consistent, well justified, and clearly stated *a priori* inclusion criteria” of studies to be

reviewed. Primary studies should be “germane to the issue at hand, should be based on a study design that minimises bias, and should have external validity”. The latter requires outcome variables that have some ‘real life’ significance rather than “extremely brief laboratory studies or other highly artificial experiments” (*ibid*).

6.10 More recently Slavin and Fashola (1998) have presented a best evidence synthesis of “proven and promising programs for America’s schools”, which uses this rather pragmatic notion of research synthesis. Some studies are included in this review even though Slavin and Fashola had reservations about some aspects of the primary studies in question. They note, for instance, that the comparison groups used in Mehan *et al*’s (1996) AVID project may be susceptible to bias, yet they conclude that “the college enrollment rates for AVID are impressive, and the program has a good track record in serving students throughout the United States. The Mehan *et al* study provides good qualitative evidence from case studies, interviews with students and teachers, and ethnographic research, of *why* and *how* the AVID programme succeeds, and has limitations. For these reasons, say Slavin and Fashola, this study is “worthy of consideration by other schools serving many students placed at risk” (Slavin and Fashola, 1998:87).

7.0 What Relevance Does All This Have for Government Research and Evaluation?

7.1 Evidence-based principles are at the heart of the Government's reform agenda for better policy making and policy implementation. "What matters is what works" is a repeated theme of government policy documents and Ministerial statements. Consequently, it is essential that Government departments and agencies have ways of accessing, harnessing and using the best available research evidence for effective policy making.

7.2 One of the frequent criticisms of systematic reviews for Government purposes is that they take a long time to complete (between six months and one year), and that potential Government users of reviews require evidence more rapidly. Establishing an evidence-base in any subject does take time, and building up a body of sound evidence is a lengthy process. Users of research and evaluation evidence often need quicker access to what the existing evidence is telling them, and what gaps remain in the research evidence on some topic or question.

7.3 *Rapid Evidence Assessments*

To this end, Rapid Evidence Assessments are being developed for use in public policy research and evaluation. Rapid Evidence Assessments are appraisals of existing evidence that sit somewhere between *the equivalent* of Health Technology Assessments (HTAs) and fully developed systematic reviews in the field of health care.

7.4 HTAs are descriptive rather than analytical abstracts of healthcare interventions that have not been critically appraised and fully evaluated according to systematic review procedures. Nonetheless, they include “evidence of clinical outcomes relative to no treatment and/or the best existing treatment for the condition in question, including undesirable side-effects and, (for chronic conditions) effects of stopping treatment” (NHS Executive, 1999). In addition, HTAs include estimates of:

- the impact on quality and length of life;
- estimates of the average health improvement per treatment initiated;
- net NHS costs associated with this health gain;
- other (non-NHS) costs and savings caused by the intervention;
- any significant differences between patients and sub-groups of the population;
- the expected total impact on NHS resources (including manpower resources).

HTAs typically take between 8 and 12 weeks to assemble.

7.5 Whilst other areas of policy and practice differ in some respects from health care there are parallels that are worth developing in terms of generating structured appraisals of what works, how, for whom, with what potential negative effects, and at what costs and benefits. Rapid Evidence Assessments will collate descriptive outlines of the available evidence on a topic, critically appraise them (including an economic appraisal), sift out studies of poor quality, and will provide an overview

of what that evidence is telling us, and what is missing from it. They will be based on fairly comprehensive electronic searches of appropriate databases, and some searching of print materials, but not the exhaustive database searching, hand searching of journals and textbooks, or searches of the grey literature that go into systematic reviews.

7.6 It is anticipated that Rapid Evidence Assessment will be completed and available in less than 8-12 weeks, though this will depend on the topic under investigation, the available evidence, and the available resources to review, appraise and summarise the evidence. Rapid Evidence Assessments will carry a caveat that their conclusions may be subject to revision once the more systematic and comprehensive review of the evidence has been completed. This is consistent with the important principle that systematic reviews are only as good as their most recent updating and revision allows.

8.0 Where Can I Find Help with Systematic Reviews and Harnessing Existing Evidence?

8.1 There are a number of academic and government agencies that provide advice, guidance and specialist expertise on how to develop, and use, systematic reviews of research evidence. Some of these agencies undertake the preparation and dissemination of systematic reviews and other types of research synthesis.

8.2 The Campbell Collaboration (<http://www.campbellcollaboration.org>) is an

international network of social scientists and policy analysts that prepares, maintains and disseminates systematic reviews of the effectiveness of interventions in education, crime and justice, and social welfare. It also provides methodological guidance and some training on how to undertake systematic reviews, and quality assurance procedures for generating valid and reliable reviews. Research and evaluation groups from around the world contribute to the Campbell Collaboration.

8.3 The Cochrane Collaboration (<http://www.cochrane.co.uk>) is the forerunner of the Campbell Collaboration and prepares, maintains and disseminates systematic reviews of the effects of interventions in health care. The Cochrane Collaboration has an impressive electronic library of systematic reviews in over 50 areas of medicine and health care. It also has nine methods groups and provides informative guidance on the methodology of systematic reviews. The Cochrane Reviewers' Handbook (available *via* the above website address) is a valuable source of guidance on how to undertake and appraise systematic reviews

8.4 The Department for Education and Skills (DfES) has commissioned a Centre for Evidence-Informed Policy and Practice in Education (the EPPI-Centre), which is located at the Institute of Education at the University of London (<http://eppi.ioe.ac.uk>). The EPPI Centre undertakes and commissions systematic reviews in education, and is developing methods for undertaking systematic reviews of social science and public policy research.

8.5 The Economic and Social Research Council (ESRC) has established an Evidence Network, which consists of a Centre for Evidence-Based Policy and Practice at Queen Mary College, London and seven evidence ‘nodes’. The Centre for EBPP is also developing the methodology of systematic reviews in the social sciences and public policy field, and is establishing a database of high quality reviews. Further details are available at <http://www.evidencenetwork.org>.

8.6 There are some very useful textbooks and handbooks on systematic reviews and research synthesis. Sources which deserve special mention are:

Cochrane Reviewers’ Handbook, 2002

<http://www.cochrane.dk/cochrane/handbook/handbook/htm>

Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Light, R.J., Louis, T.A. and Mosteller, F., 1992

Meta-Analysis for Explanation, New York, Russell Sage Foundation.

Cooper, H. and Hedges, L.V. (eds), 1994

The Handbook of Research Synthesis, New York, Russell Sage Foundation.

Egger, M., Davey Smith, G. and Altman, D.G. (eds), 2001

Systematic Reviews in Health Care: Meta-Analysis in Context, London, BMJ Publishing Group.

Slavin, R.E. and Fashola, O. S., 1998

Show Me the Evidence! : Proven and Promising Programs for American Schools, Thousand Oaks, California, Corwin Press.

8.7 An Analysts’ Checklist for Undertaking a Systematic Review is presented at Annexe 1 below.

8.8 A Policy Makers’ Checklist for Using a Systematic Review is presented at Annexe 2 below.

9.0 Conclusion

9.1 There is a growing recognition of the potential of systematic reviews and other types of research synthesis for policy evaluation and analysis. Systematic reviews provide a powerful way of harnessing existing evidence that is valid, reliable and transparent. They differ from traditional narrative reviews and other types of literature review in that they use exhaustive methods for searching evidence, critically appraise that evidence according to explicit criteria, and identify the implications for policy and practice only from research that has passed high quality control procedures. Systematic reviews are not a panacea, nor are they a substitute for sound judgement and expert decision making. Rather, they provide one means of establishing a sound empirical research base upon which the judgements and expertise of decision makers can be made.

Annexe 1

Analysts' Checklist for Undertaking a Systematic Review

This checklist should only be used in conjunction with one or more of the existing guides to systematic reviews mentioned in paragraph 8.6 of these guidance notes. Further advice and guidance can be obtained from the Policy Evaluation Division of the Cabinet Office Strategy Unit (phil.davies@cabinet-office.x.gsi.gov.uk)

Analysts proposing to undertake a systematic review for the first time are also advised to take a structured course on the topic, such as the Cabinet Office course on 'Harnessing the Evidence: Systematic Reviews and Meta-Analysis'.

1. Formulating an Answerable Question

Does the central question of the review clearly address the following points?

- The policy *intervention* for which evidence is sought
- The *population* or *sub-groups* that the policy is expected to effect
- The *outcomes* that the policy intervention is expected to achieve

2. Searching For Relevant Studies

Have the following steps of a search strategy been planned?:

- The searching of appropriate *electronic/internet* sources?
- The searching of appropriate *print* sources (e.g. journals, textbooks, research reports)?
- The *hand searching* of appropriate print sources?
- Searching of the 'grey' (i.e. unpublished) literature?

3. Critically Appraising Studies Found

How will the existing literature be sifted for quality and validity?

- The *appropriateness* of the questions, populations and outcomes addressed
- Evidence of *selection bias* in the primary studies
- Evidence of *performance bias* in the primary studies
- Evidence of *attrition bias* in the primary studies
- Evidence of *detection bias* in the primary studies

- What criteria will be used for *including and excluding* primary studies

4. Extracting Data From Included Studies

Has a strategy been planned for extracting data from the included studies?

- A data collection form recording how, and why, data were extracted from included studies
- Information about the characteristic of included studies
- Verification of study eligibility for the review
- Details of study characteristics
- Details of study methods
- Details of study participants (populations and sub-groups)
- Details of study interventions
- Details of study outcomes and findings
- Reliability check for data collection/extraction

5. Analysing and Presenting The Findings

- What comparisons should be made (e.g. by interventions studied, participants included, outcomes measured)?
- What study results are needed for each comparison?
- What assessments of validity are to be used in the analysis?
- Is any other data or information needed from authors of studies included in the review?
- Do the data from different studies need to be transformed for the review's analysis?
- How is heterogeneity/homogeneity of studies to be determined?
- Is a meta-analysis of findings possible?

- What are the main findings of the review?
- What are the likely effect sizes of the proposed policy intervention, net of the counterfactual?
- What are the main caveats and/or qualifications of the findings of this review?

6. Interpreting the Findings

- What is the strength of the evidence from the review?
- How applicable are the results of the review to 'real life' policy and practice?
- What does the review say about the costs and benefits of the proposed intervention?
- What trade-offs are suggested by the review between expected benefits, harm and costs (including opportunity costs)?
- What mediating factors emerge from the review that might affect the implications for policy and practice in different contexts?

7. Summarising the Implications for Policy and Practice

- What are the 'take home' messages for policy making and/or practice?
- What are the 'take home' messages for future research in this area?

Annexe 2

Policy Makers' Checklist for Using a Systematic Review

The following are suggested questions that policy makers should ask when reading a systematic review. Further advice and guidance can be obtained from the Policy Evaluation Division of the Cabinet Office Strategy Unit (phil.davies@cabinet-office.x.gsi.gov.uk)

1. The Question You Want Answered

Has the reviewed covered the following features?:

- The policy intervention for which evidence is sought
- The population or sub-groups that the policy is expected to effect
- The outcomes that the policy intervention is expected to achieve

2. The Adequacy of the Review

Are there sufficient details in the review about the search strategy used to find relevant research evidence?: i.e.

- Were appropriate electronic/internet sources searched?
- Were appropriate print sources (e.g. journals, textbooks, research reports) searched?
- Was any hand searching of appropriate print sources used?
- Was an attempt made to identify the 'grey' (i.e. unpublished) literature

3. Critical Appraisal of Evidence

Were the following tasks undertaken in the review?

- Was the research evidence that was identified sifted and graded for quality?
- Were the inclusion and exclusion criteria of primary studies made explicit?
- Were the appropriate outcome measures included in the review?

4. Quality of Evidence Presented

- Is the research evidence presented in the review easy to understand?
- Is a full evidence table presented?
- Has the strength of the existing evidence been assessed?
- Are there any estimates of the likely effect size of the policy intervention?

- Are there any details about the contexts in which the policy is likely to be effective?
- Is there any information on the likely costs and benefits of the proposed policy?

5. Applicability of the Evidence

- Has the evidence been presented in a way that is helpful to decision making?
- What are the implications of the review for policy and practice in this area?
- Are there any mediating factors from the review that need to be taken into account?

- What are the implications of this review for future research and evaluation in this area?

6. Peer Review

- Has this systematic review been peer reviewed by independent analysts with expertise and experience in research synthesis?

- If not, do you want to get this systematic review peer reviewed?

References

Cabinet Office, 2002

‘The Quality of Qualitative Research’, Research project commissioned by the Centre for Management and Policy Studies, Cabinet Office, being carried out by the National Centre for Social Research, London, CMPS, Cabinet Office.

Cook, T.D., Cooper, H., Cordray, D.S., Hartmann, H., Light, R.J., Louis, T.A. and Mosteller, F., 1992

Meta-Analysis for Explanation, New York, Russell Sage Foundation.

Cooper, H. and Hedges, L.V. (eds), 1994

The Handbook of Research Synthesis, New York, Russell Sage Foundation.

Davies, P.T., 2000

‘Qualitative Research Methods in Evidence-Based Policy and Practice’, in Davies, H.T.O., Nutley, S.M. and Smith, P.C. (eds), *Evidence and Public Policy*, Bristol, The Policy Press (forthcoming).

Deeks, J.J., Altman, D.G. and Bradburn, M.J., 2001

‘Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis’, in Egger, M., Davey Smith, G. and Altman, D.G. (eds), Systematic Reviews in Health Care: Meta-Analysis in Context, London, BMJ Publishing Group.

Eddy, D. 1999

Doctors, Economics and Clinical Practice Guidelines: Can they Be Brought Together?, London, Office of Health Economics.

Egger, M., Davey Smith, G., and O’Rourke, K. 2001

‘Rationale, potentials, and promise of systematic reviews’, in Egger, M., Davey Smith, G. and Altman, D.G. (eds), Systematic Reviews in Health Care: Meta-Analysis in Context, London, BMJ Publishing Group.

Egger, M., Davey Smith, G. and Altman, D.G. (eds), 2001

Systematic Reviews in Health Care: Meta-Analysis in Context, London, BMJ Publishing Group.

Egger, M., Davey Smith, G., and Schneider, M., 2001

‘Systematic reviews of observational studies’, in Egger, M., Davey Smith, G. and Altman, D.G. (eds), Systematic Reviews in Health Care: Meta-Analysis in Context, London, BMJ Publishing Group.

EPPI-Centre, 2001

Review Group Manual, Version 1.1, London, Institute of Education.

- Glass, G.V., 1976
 'Primary, secondary and meta-analysis of research, Educational Researcher, **5**, 3-8.
- Glass, G.V. and Smith, M.L., 1979
 'Meta-analysis of research on class size and achievement', Educational Evaluation and Policy Analysis, **1**, 2-16.
- Glass, G.V., Cahen, L.S., Smith, M.L., and Filby, N.N. 1982
School Class Size: Research and Policy, Beverley Hills, Sage Publications,
- Hedges, L.V. 1994
 'Fixed Effects Models', in Cooper, H. and Hedges, L.V. (eds), The Handbook of Research Synthesis, New York, Russell Sage Foundation.
- Kulik, J.A. and Kulik, C-L. C. 1989
 'Meta-analysis in education', International Journal of Educational Research, **13**, 221-340.
- Mehan, H., Villanueva, T., Hubbard, L. and Lintz, A., 1996
Constructing School Success: The Consequences of Untracking Low-Achieving Students, Cambridge, Cambridge University Press.
- Morton, S., 1999
 'Systematic Reviews and Meta-Analysis', Workshop materials on Evidence-Based Health Care, University of California, San Diego, La Jolla, California, Extended Studies and Public Programs.
- Noblit, G.W. and Hare, R.D., 1988
Meta-Ethnography: Synthesizing Qualitative Studies, Newbury Park, Sage Publications.
- NHS Executive, 1999
 'Faster Access to Modern Treatment: How NICE Appraisal Will Work', Leeds, National Health Service Executive.
- Oakley, A., Gough, D., Harden, A., (personal communication)
 Quality Standards for Systematic Synthesis of Qualitative Research', Research project, EPPI-Centre, Institute of Education, University of London.
- Raudenbusch, S.W. 1994
 'Random Effects Models', in Cooper, H. and Hedges, L.V. (eds), 1994The Handbook of Research Synthesis, New York, Russell Sage Foundation.
- Slavin, R. E. 1984

'Meta-analysis in education: How has it been used?', Educational Researcher, **13**, 6-15.

Slavin, R.E. 1986

'Best evidence synthesis: An alternative to meta-analysis and traditional reviews, Educational Researcher, **15**, 5-11.

Slavin, R.E. and Fashola, O. S., 1998

Show Me the Evidence! : Proven and Promising Programs for American Schools, Thousand Oaks, California, Corwin Press.

Smith, M.L. and Glass, G.V., 1980

'Meta-analysis of research on class size and its relationship to attitudes and instruction', American Educational Research Journal. **17**, 419-433.

Smith, M.L., Glass, G.V. and Miller, T.I., 1980

The Benefits of Psychotherapy, Baltimore, Johns Hopkins University Press.

Petrosino, A.J., Boruch, R. Rounding, C., McDonald, S., and Chalmers, I., 2002

A Social, Psychological, Educational & Criminological Trials Register (SPECTR), Philadelphia, Campbell Collaboration (<http://campbell.gse.upenn.edu>)