UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES

School of Education

Assessing quality in systematic reviews of the effectiveness of health promotion: Areas of consensus and dissension

by Jonathan Paul Shepherd

BA (Hons), MPhil

Thesis for the degree of Doctor of Philosophy

June 2009

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES,
SCHOOL OF EDUCATION

Doctor of Philosophy

ASSESSING QUALITY IN SYSTEMATIC REVIEWS OF THE EFFECTIVENESS OF HEALTH PROMOTION: AREAS OF CONSENSUS AND DISSENSION

By Jonathan Paul Shepherd

Systematic reviews have played an increasingly important role in health promotion in recent years. Yet there are debates about how they should be conducted, particularly about how the quality of evidence should be assessed. The aim of this research was to assess current approaches to, and general views on, the use of quality assessment in systematic reviews of effectiveness in health promotion, and to identify areas of consensus and dissension around the choice of techniques, methods and criteria employed.

There were two stages of data collection. The first was a structured mapping of a random sample of 30 systematic reviews of the effectiveness of health promotion to identify and explain trends and themes in methods and approaches to quality assessment. During the second stage semi-structured interviews were conducted with a purposive sample of 17 systematic reviewers who had conducted at least one review of a health promotion topic, to investigate some of these trends and approaches in greater detail.

The mapping found that the majority of systematic reviews had assessed the quality of the included studies, to varying degrees. However, procedures were not always explicitly reported or consistent. There was some degree of consensus over criteria, with experimental evaluation methods commonly favoured. Most frequently used quality assessment criteria included participant attrition, the validity and reliability of data collection and analysis methods, and adequacy of sample sizes. External validity was commonly assessed, primarily in terms of generalisability and replicability, but less so in terms of intervention quality.

The interviews revealed some of the barriers to effective systematic reviewing, including: lack of time and resources, complexity of some health promotion interventions, inclusion of observational evaluation designs, and poor reporting of primary studies. Systematic reviewing was commonly done in small teams, mostly comprising academics, sometimes with practitioners. Interviewees learned systematic review skills through a combination of training, support from colleagues and mentors, literature and a strong emphasis on hands-on practical learning. Subjective judgement was often required, contra to the popular belief that systematic reviews are wholly objective.

The overall conclusions of this study are that systematic reviewing in health promotion is often challenging due the complexity of interventions and evaluation designs. This places additional demands on reviewers in terms of knowledge and skills required, often exacerbated by finite time scales and limited funding. Initiatives are in place to foster shared ways of working, although the extent to which complete consensus is achievable in a multi- disciplinary area such as health promotion is questionable.

Table of contents

List of tables

List of figures

Author's declaration

I, Jonathan Shepherd, declare that the thesis entitled 'Assessing quality in systematic reviews of the effectiveness of health promotion: Areas of consensus and dissension' and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed:

Date: 11th June 2009

Acknowledgements

I would like to thank my supervisor, Professor Katherine Weare, for her sound advice and friendship throughout.

Thanks also go to my employer, Southampton Health Technology Assessments Centre (SHTAC), Wessex Institute, University of Southampton, for financial support throughout the project.

I am indebted to all of the interviewees who generously gave up their time for this research.

Finally, thanks go to Mikey for proof reading and for all his support, and to my dear Sam who was always such a good friend and who I will miss greatly.

Abbreviations

CDSR	Cochrane Database of Systematic Reviews
CONSORT	Consolidated Standards of Reporting Trials
CRD	Centre for Reviews and Dissemination
DARE	Database of Abstracts of Reviews of Effectiveness
DoPHER	Database of Public Health Effectiveness Reviews
DPP	Disease Prevention Panel
DVT	Deep Vein Thrombosis
EPPI - Centre	Evidence for Policy and Practice Information and Co-ordinating Centre
ESRC	Economic and Social Research Council
HTA	Health Technology Assessment
LMWH	Low Molecular Weight Heparin
QUOROM	Quality of Reporting of Meta-Analysis
NICE	National Institute for Health and Clinical Excellence
NIHR	National Institute for Health Research
RCT	Randomised Controlled Trial
TREND	Transparent Reporting of Evaluations with Nonrandomized Designs

Introduction

The position of the author

This topic of this thesis was chosen because of its relevance to my area of research; the production of systematic reviews of the effectiveness of health care and health promotion. I first began working in health promotion research in the School of Education at the University of Southampton in 1995, evaluating the effectiveness of a peer-led HIV prevention intervention in the gay community in Southampton. This two-year project provided me with a grounding in the theory and practice of evaluation and health promotion. At this time the debates about the use of randomised controlled trials to evaluate health interventions were gathering pace, and the project team were strongly encouraged by the funding body to use a controlled design. We adopted this design and it gave me experience of the challenges of applying the (quasi) experimental method, often used in settings with a high degree of control over the research process, in a dynamic and complex community setting.

This experience stimulated my interest in the concepts of study design, reliability and validity. In 1997 I became involved in my first systematic review on the effectiveness of interventions to promote sexual health, and in 1998 took up a post at the Wessex Institute for Health Research and Development within the University (where I remain today) conducting systematic reviews of the effectiveness of health care interventions initially for the (former) South and West National Health Service (NHS) Development and Evaluation Committee (DEC), and latterly for the National Institute of Health and Clinical Excellence (NICE). During this time I became active within the Cochrane Collaboration, an international network of individuals dedicated to producing high quality systematic reviews of effectiveness of health care. I became a co-ordinator of the Cochrane Health Promotion and Public Health Field, and took an 18 month secondment to the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre), Institute of Education, University of London where I returned to researching the effectiveness of health promotion. This experience took me into the heart of the debates about evidence for effective health promotion, enabling me to establish a network of colleagues with whom I continue to collaborate on various research projects. This network has proved invaluable in conducting this research.

As a systematic reviewer I became interested in assessing the methodological quality of the studies included in the review. My earlier experience of primary evaluation gave me an understanding of the realities of striving to produce rigorous evidence of effectiveness in the

complex 'real world' setting, the environment in which most health promotion takes place. As a reviewer, I was charged with the responsibility of appraising the very kind of studies I had already been involved in.

Interestingly, my earlier evaluation study was later included in a systematic review of the effectiveness of sexual health interventions for men who have sex with men (conducted by colleagues at the EPPI-Centre) (Rees *et al*, 2004a). Although I was a part of the review team, I was not directly involved in the quality assessment of the included studies. To my relief the study was considered to be of sufficient quality to qualify for the sub-set of studies used to support conclusions and recommendations. I found myself in the unusual position of being involved in a systematic review of evaluation research with personal insight into the realities of the evaluation process. It raised questions for me about whether experience of the challenges of evaluation biased me as a systematic reviewer into perhaps being too lenient when appraising studies, potentially giving too much benefit of the doubt, with the risk of underplaying their methodological shortcomings. Should a reviewer strive for objectivity at the expense of empathy and tolerance? I then started thinking about wider questions, such as what background, training and experience does a systematic reviewer require to enable them to appraise evaluation research effectively? On what grounds should we be judging the quality of evaluation research? Is there any consensus on approaches and techniques for appraising evidence? It is questions such as these which prompted me to conduct this research.

The primary reason for undertaking a *systematic* review of the evidence, as opposed to a what could be described as a literature review, is to be objective and transparent in the identification, selection, appraisal and analysis of studies in order to minimise bias in conclusions about effectiveness. Yet it is important to recognise my own biases. Whilst I believe in the importance of evidence based health and the need for systematic reviews, I acknowledge that the methodology remains relatively elementary and has been open to criticism. With this in mind, it is important to take a critical view of systematic reviews themselves and identify key areas where the methodology could be improved. I have chosen the appraisal of evidence as one particular area of systematic reviews where there are unanswered questions about methodology. My experience of being on both sides of the primary and secondary evaluation fence, and of reviewing the effectiveness of both health promotion and health care, has provided me with (I hope) a balanced perspective for this research. This is the context in which this thesis has been written.

Chapter 1 - Rationale for the research

Chapter outline

The aim of this chapter is to provide an outline of, and justification for, the issues to be investigated in this thesis. It begins by introducing the concept of evidence-based health and systematic reviews, and discusses the strengths and weaknesses of this approach to evidence. The process of quality assessment within a systematic review is then discussed in detail and key areas for further investigation (in terms of research objectives), are identified. These include: documenting the extent to which systematic reviews assess quality, the need to establish whether there is any consensus on which dimensions of internal and external validity should be assessed in systematic reviews, assessing at what stage of a review quality assessment is commonly undertaken, how judgement of quality is used in the analysis of results and presentation of conclusions, who might be best placed to assess quality, and what training and support they might need. The chapter concludes with the setting of aims and a summary of research objectives and an outline of the methodology.

The research objectives proposed throughout this chapter are based on a thorough critical review of the literature in evidence-based health promotion, and also by a round of agenda-setting interviews with a small number of experts in the area (the methodology of these interviews is described in Chapter 2 and Appendix 1).

The aims of this study are:

1) To assess current approaches to, and general views on, systematic reviews of the effectiveness of health promotion; specifically the process of quality assessment of included evidence.
2) To identify areas of consensus and dissension around the choice of techniques, methods and criteria employed.

1.1 The evolution of evidence-based health and systematic reviews

Over recent years there has been a drive toward improving the methods used to evaluate the effectiveness of interventions in the field of health and medicine, commonly known as 'evidence-based health' (sometimes also referred to as evidence-informed health). The primary aim is to ensure that policy and practice is influenced by sound and reliable evidence, maximising benefits and minimising potential harm. One of the founding fathers of the movement was Archie Cochrane, an epidemiologist who in the early 1970s, as Tones and

Tilford (2001: 35) describe, 'began to ask rather awkward questions about the effectiveness of many routinely accepted medical procedures'. Cochrane questioned why there was no critical summary of all relevant RCTs in health (Cochrane, 1979). As will become apparent in this chapter, Cochrane's vision was to later become the foundation for the Cochrane Collaboration, an organisation whose aim is to promote evidence-based health worldwide.

1.1.1 The hierarchy of evidence

Explicit in the principles of evidence-based health is the view that certain evaluation designs (e.g. experimental designs, in which one group of people receive an intervention whilst another group receive an alternative or nothing at all) are at less risk of bias than non-experimental studies (hereafter referred to as observational studies). It is suggested that, where feasible, RCTs provide the most rigorous evidence upon which conclusions regarding efficacy can be based (Altman and Bland, 1999; Kleijnen *et al*., 1997; Maynard and Chalmers, 1997).

The basic principle behind the RCT is quite simple. Intervention recipients are randomly allocated, by chance, to receive either an experimental intervention, an alternative intervention, or no intervention at all. The benefit of randomisation is that it achieves an even distribution of participants to the intervention and comparison groups in terms of characteristics known and unknown to influence outcomes (e.g. age, sex, education, or health-related attributes) (Schulz and Grimes, 2002). As will be explained later in Section 1.5, this protects against 'selection bias'. When groups are adequately matched on these characteristics the investigator can attribute with greater confidence the observed changes in outcomes to the intervention, rather than pre-existing differences between them.

Experimental designs, specifically RCTs, are therefore prioritised to support decision making in health care in a hierarchy of evidence (Table 1). The purpose of the hierarchy is to facilitate evidence-based decision making. For example, a policy maker looking for evidence of effectiveness to underpin a proposed strategy may use it to help prioritise which evidence to use. Prioritisation may be important where the volume of literature is high and busy schedules do not allow much time for reading. Thus, when faced with a vast number of evaluation reports spanning all levels of the hierarchy, the policy maker may choose only to read reports of the experimental evaluations, for instance, given that they may be more reliable than the observational studies.

The basis of the hierarchy is not arbitrary. It is based on the results of empirical methodological studies which have demonstrated that evaluation designs lower down the hierarchy (and even

Table 1 – Hierarchy of evidence

Evaluation design hierarchy	
Level	*Description*
1	Experimental studies (e.g. RCT with concealed allocation of participants)
2	Quasi- Experimental studies (e.g. non-randomised controlled studies; before and after study; interrupted time series)
3	Observational studies (e.g. cohort studies; case control studies; case series)

Adapted from: Centre for Reviews and Dissemination (2009)

those at the top if not conducted appropriately) are more likely to over-estimate intervention effects (Guyatt *et al,* 2000; Schulz *et al*, 1995). The findings of more recent studies, however, have complicated the picture. They have found that in some cases effects between randomised and non-randomised studies are similar, and in other cases different, with no consistent pattern in effects (Deeks *et al*, 2003; Oliver *et al*, 2008). Nonetheless, on theoretical grounds RCTs are still recommended as being the most rigorous evaluation design (Oliver *et al*, 2008).

1.1.2 The use of systematic reviews

Accompanying the drive for higher standards in the evaluation of health interventions has been a proliferation of systematic reviews of the literature. These reviews draw together the results of primary evaluations in a manageable summary, the strengths and weaknesses of which are discussed in the next section (Section 1.2)

An infrastructure for the promotion of evidence-based health has been evolving since the early 1990s, with the establishment of various organisations, perhaps the most prominent being the Cochrane Collaboration. As mentioned earlier, the Collaboration is an international network of individuals and groups whose task is dedicated to preparing, maintaining and promoting the accessibility of systematic reviews of the effects of health care interventions (Clarke, 2006; Higgins and Green, 2008). The English Department of Health has also shown commitment to the aim of establishing a knowledge-based health service (McGuire, 2006). Health strategies such as 'Choosing Health', the White Paper for public health in England (Department of Health, 2004), stress the need for sound evidence to underpin policies, to enable crucial targets to be met and to increase public accountability.

This is not just empty rhetoric. There are a number of health priorities for which sound evidence is needed to underpin effective interventions, including: coronary heart disease (Department of Health, 2009), obesity (Cross-Government Obesity Unit, 2008), depression and

suicide/self-harm (particularly in young men) (Department of Health, 2002a), teenage pregnancy (Teenage Pregnancy Independent Advisory Group, 2008) and sexually transmitted infections, particularly among young people (Health Protection Agency, 2008). The National Strategy for HIV/AIDS and Sexual Health, for example, described the evidence base in the area as being "dispersed and unsystematic" (Department of Health, 2001: 17) and the accompanying implementation action plan stressed the need for systematic reviews of the literature to support the strategy (Department of Health, 2002b).

The Department of Health has therefore funded a number of organisations including the UK Cochrane Centre in Oxford, the Centre for Reviews and Dissemination (CRD) at the University of York (Sowden and Glanville, 2006), the National Institute for Health and Clinical Excellence (NICE) (incorporating the former Health Development Agency) (Kelly, 2005; Littlejohns, 2006), the Evidence for Policy and Practice Information and Coordinating Centre (EPPI-Centre) at the Institute of Education, London (Oakley *et al*, 2005; Oliver *et al*, 2006), and The National Institute for Health Research (NIHR) Health Technology Assessment Programme (HTA). Collectively they are responsible for building the evidence base, through activities such as commissioning and production of primary and secondary evaluative research, and publication of guidance to the National Health Service. Notably, in 2004 the HTA programme initiated the Disease Prevention Panel specifically to prioritise topics for primary research (i.e. RCTs) and secondary research (i.e. systematic reviews) in the area of health promotion for funding. It is therefore evident that there has been an increase in demand for high quality evidence of effectiveness, including health promotion.

The drive for evidence-based policy and practice has also gathered pace in other disciplines. For example, in 2000 the English Department for Education and Skills (DFES) (now The Department for Children, Schools and Families and the Department for Innovation, Universities and Skills) funded the EPPI-Centre to facilitate the production of systematic reviews of effectiveness in education. The Campbell Collaboration, a sibling of the Cochrane Collaboration, was established in 2000 with the remit of promoting an evidence-based approach in education, criminology, psychology, social work and social policy (Boruch *et al*, 2004). The Economic and Social Research Council (ESRC) has also funded the establishment of an 'Evidence Network', comprising a number of collaborating academic centres which act as focus points for evidence-based policy and practice research (Petticrew *et al*, 2006).

It is evident that a substantial amount has been invested in evidence-based health, particularly in the UK. This is set against a backdrop of demands for greater accountability, and calls for public policy to be influenced by the rigorous evidence. Systematic reviews of effectiveness have

become one of the main tools of evidence-based health, yet there is considerable debate about the most appropriate methods for conducting them, particularly in relation to what kind of evidence is included and how it is appraised. The next section introduces the concept of systematic reviewing, critically discusses their strengths and weaknesses and proposes unresolved issues that will be investigated in this thesis.

1.2 A critical appraisal of systematic reviews

1.2.1 Defining systematic reviews

As mentioned earlier, in recent years there has been a surge in publication of systematic reviews of the effectiveness of health interventions. The practice of combining studies, however, is not new. There are examples of meta-analyses (defined as systematic reviews that use statistical methods to pool studies to produce an overall quantitative estimate of effect) in the field of education and psychology extending back to the 1970s (see Fitz-Gibbon, 1985 and Oakley, 2000 for a discussion of these). It is largely in the 1980s and 1990s that the practice has gathered momentum in the field of health.

Clarke and Oxman (2001: 27) capture the main characteristics of a systematic review in their definition:

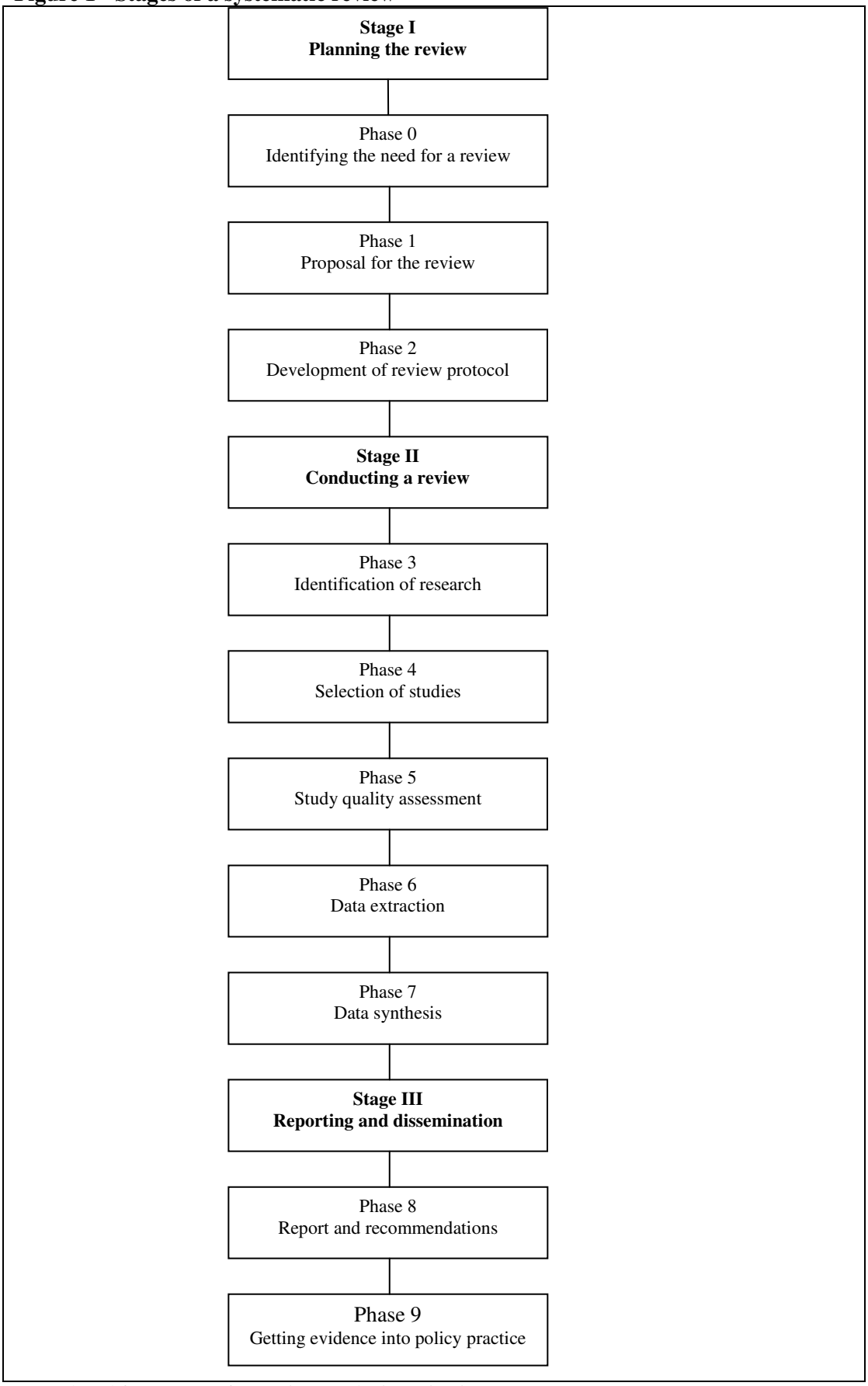
"A review of a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review. Statistical methods (meta-analysis) may or may not be used to analyse and summarise the results of the included studies"

Although the precise characteristics of reviews vary, there are usually a number of key stages (Figure 1) including: formulation of the review question, development of inclusion and exclusion criteria for studies, writing and publishing a protocol (i.e. a proposal), searching for and retrieving reports of studies, applying inclusion and exclusion criteria to retrieved reports, extracting data and critically appraising the included studies, and combining them in a synthesis to form conclusions and recommendations for policy and practice.

1.2.2 Strengths of systematic reviews

One of the key benefits of a systematic review (and one which distinguishes it from what might be called a non-systematic review, also sometimes referred to as a 'literature review') is that, if conducted correctly, it should represent a comprehensive and sound overview of the literature in a given area.

Figure 1 - Stages of a systematic review



(reproduced from Centre for Reviews and Dissemination (2001))

Peersman *et al* (2001: 26) discuss this in relation to the question 'What makes a review systematic?':

"The methods used in conducting a systematic review aim to limit both systematic errors (bias) and errors that occur by chance (random errors). These methods are explicitly reported so that others can assess the integrity of the review process, and hence, the validity of the review"

For example, comprehensive literature searching is regarded as one of the key distinguishing factors between systematic and non-systematic reviews (Higgins and Green, 2008). The aim is to locate as much relevant literature as possible to ensure the results and conclusions of the review are based on all of the available evidence. This will include searching for studies which may be difficult to access, such as unpublished studies, or those published in languages other than English. Without exhaustive searching, the conclusions of the review may be biased in favour of published studies. This is particularly important as published studies tend to be more likely to conclude that interventions are effective (Dickersin *et al,* 1987; Dickersin, 1997). It is therefore important that all the stages of a review, as in any piece of research, are conducted in a sound and methodical way.

A second strength of the systematic review lies in its ability to define the impact of an intervention more precisely than a single primary evaluation. There are two main limitations in relying on the answer produced by a single evaluation of an intervention. The first is lack of sufficient precision with which the effect of an intervention can be detected, the second is difficulty in generalising the results of a trial beyond the specific study participants and intervention tested (Thompson and Pocock, 1991). In other words, the results of one evaluation may be spurious or due to chance and it is only by looking across a number of evaluations that we can be confident in our conclusions about the effectiveness of interventions. In certain circumstances statistically combining the results of a number of averaged sized studies will boost the total sample size and potentially increase the ability to precisely detect an effect of the intervention if one exists (Dickersin and Berlin, 1992). This may be particularly useful in health promotion where many evaluations may not have sufficiently large enough sample sizes to show an effect (O'Leary *et al*, 1997).

A systematic review, therefore, is not just a passive summary of the extant literature in a given area. Rather, its added value is the clarity it provides about the precise impact of a particular health intervention. It can thus be acknowledged as a piece of research in its own right.

The fact that a more definitive answer can be achieved by statistically pooling together the results of a number of evaluations underscores the need for comprehensive literature searching to identify as much of the literature as possible. This is illustrated by a much quoted example of a systematic review of corticosteroids to prevent complications from premature birth. Early clinical trials of corticosteroids to prevent complications indicated that generally the treatment was associated with only moderate benefit. However, it was only when all the trials were first combined in 1989 in a systematic review (see Crowley, 2000), that the extent of the benefits, a reduction in the risk of babies dying from the complications of immaturity, became apparent (NB. The graphical representation of this meta-analysis, or 'forest plot', is used by the Cochrane Collaboration as their logo). In short, the drug was far more effective than previously thought. It is suggested that had the review been conducted earlier, the drug would have been more widely used and thousands of premature babies would not have suffered and died unnecessarily (Cochrane Collaboration, 2004).

Similar discrepancies in findings from single evaluations compared to pooling trials in systematic reviews are apparent within health promotion. In a review of smoking cessation in pregnancy (Lumley *et al*, 2000), only 12 out of 34 trials showed a statistically significant effect on women reporting giving up. However, pooling the data from all trials showed a clear substantial benefit of intervention. Systematic reviews can also show the opposite. In a review of the effects of commercial breastfeeding promotional materials for new mothers at hospital discharge, only two of the nine trials showed a negative impact on exclusive breastfeeding (Donnelly *et al*, 2000). However, when the data from trials was pooled, a significant detrimental effect was detected, namely a reduction in the period of exclusive breastfeeding. Therefore, one of the advantages of systematic reviews, particularly those in which statistical techniques such as meta-analyses are employed, is the power to detect not only benefit, but also harm.

A third advantage of systematic reviews is their ability summarise vast sums of literature into a manageable overview, and therefore be a valuable resource for busy policy makers, practitioners, researchers and the public (including people sometimes referred to as 'health care consumers' on account of their use of health services), who rarely have the time or the resources to identify, appraise, and digest all of the evidence themselves. In some areas, where a great deal of evaluation literature is available, systematic reviews are a necessity to aid decision making (e.g. in the prevention of HIV/AIDS, see Shepherd and Harden, 2003).

1.2.3 Potential limitations of systematic reviews

One criticism is that, traditionally, systematic reviews have answered narrowly-focused questions. Martyn Hammersley, for example, argues that the practice of systematic reviewing is rooted in a positivist paradigm and is therefore subject to the same criticisms that have been made of positivism for many years (Hammersley, 2001). These include the adoption of a narrowly-focused quantitative and reductionist perspective on evidence which fails to do justice to the inherent complexity of practice situations. In this sense evidence-based health denies methodological pluralism, prioritising quantitative evidence to answer questions such as: Does it work? Broader questions may be overlooked, such as: For whom does it work? And is it appropriate?

The reason for the narrow focus is partly because many reviews have tended to prioritise the inclusion of data from experimental evaluations, which themselves pursue limited questions:

"Reviews of effectiveness which limit their scope to studies with experimental or quasi-experimental designs are of value, but in the end may be able to draw only very limited conclusions about specific types of intervention"
(Whitehead, 1996: 1).

Tones (2000: 228) echoes this by commenting with reference to the RCT:

"We might learn that a given intervention has been successful or has failed but we would normally not know why. Without illumination provided by alternative research designs it is impossible to develop and improve the programme"

However, an increasing number of experimental evaluations are conducting integral process evaluation in order to assess, amongst other things, how the intervention was implemented, and what the participants thought of it (Oakley *et al*, 2006; Stephenson *et al*, 2004). Examples of systematic reviews which have included process evaluations as well as outcome evaluations (thus attempting to answer 'Does it work?' alongside 'Why does it work? and 'For whom?') have been published, including my own work (Brunton *et al*, 2005; Rees *et al*, 2006; Shepherd *et al*, 2006a). The critique that systematic reviews answer a limited range of questions and do not engage with qualitative research is not necessarily justified. However, as there are still few published examples of reviews embracing broader questions it is necessary to discuss and debate the issue further to identify any barriers to their production.

Another criticism is that systematic reviews have relied on experimental evidence to answer questions of effectiveness at the expense of other evaluation designs (Speller *et al*, 1997). Some suggest the utility of systematic reviews to be limited because they exclude interventions which have not been evaluated to what is considered to be a high standard, but which may nevertheless be of interest to users of reviews (Ogilvie *et al*, 2005; White, 2001). This is exacerbated by the fact that, as will be discussed later, the more complex interventions which arguably might be more effective as they tackle a number of different determinants of health, are often not amenable to experimental evaluation and therefore tend not to be included in reviews.

Practitioners are therefore often presented with systematic reviews which report on the effectiveness of only a small proportion of health promotion activity (Ogilvie *et al,* 2005; Tilford, 2000). For example, in a review of the effectiveness of promoting physical activity with young people (Rees *et al*, 2001) of the 186 relevant studies screened for inclusion, only four were deemed to be of sufficient quality to support the review's conclusions. However, the alternative option of relaxing the criteria to include studies lower down the hierarchy of evidence could compromise the quality of the review, leading to conclusions and recommendations based on research with higher risk of bias. Given the potential for some interventions to do harm, as mentioned earlier, this could have drastic consequences for health.

One emerging approach which has been devised in an attempt to address many of these criticisms is 'realist synthesis' (Pawson, 2006a; Pawson *et al*, 2005). Based on concepts of realism, it uses theory to attempt to elucidate in detail 'what works for whom, in what circumstances, in what respects and how' (Pawson *et al*, 2005: 21). In this respect it attempts to answer broader questions about the nature of social, particularly complex, interventions, drawing on multiple forms of evidence (e.g. evaluation, action research, surveys, documentary analysis). It would seem then, that this form of evidence synthesis can compensate for some of the criticisms made of systematic reviews to date. Yet, as acknowledged by Pawson *et al* (2005), the method is subject to certain limitations. It is unlikely to be able to provide the definitive answer that many decision makers need about whether or not an intervention works, and as yet there are few published examples of its application, particularly in health. It also requires a high level of skill and research experience to conduct, raising questions about its practicability as a routine method of evidence synthesis. A key issue for investigation therefore is how the limitations of systematic reviews can feasibly be overcome through promising approaches such as realist synthesis.

To recap, this section has discussed some of the strengths and limitations of systematic reviews. It is acknowledged that they are comprehensive in their searching for literature, show the bigger

picture through combining all available studies and can provide users with a succinct overview of the evidence base. However, they have been criticised for employing inclusion criteria that are too stringent, often resulting in a review that is based on only a fraction of the extant literature in a given topic area.

What do systematic reviewers have to say in reply to these criticisms? What do they see as being the strengths and weaknesses of the reviews they produce? Do they themselves consider there to be other strengths or weaknesses? There have been few publications in which systematic reviewers have reflected and discussed these issues, and what little there is has been in the field of education (for example, see Oakley, 2003; Nind, 2006) or social policy (Wallace *et al*, 2006) rather than health promotion, underlining the importance of this research. An objective for this research is therefore:

Research objective:

1. To assess current views on the strengths and weaknesses of systematic reviews of health promotion

Another area that has received little investigation is the challenges that systematic reviewers in health promotion have faced. Such challenges might be political, practical or philosophical. Yet once again, there seems to have been few published accounts by systematic reviewers of obstacles they have encountered, and strategies they have used to overcome them. One of the few examples is by Nind (2006) who reflected on her experiences of conducting a systematic review of pedagogical approaches for special educational needs. She noted that although systematic reviews are explicit and transparent about their methods, there is a lack of reporting about the challenges that arise, and the sometimes problematic decisions that have to be made during the review process:

'Needless to say for every transparent account of the process of applying inclusion and exclusion criteria and so on in systematic review there is another story of pragmatic decision-making and subtle judgement. Perhaps all we need is some honesty about this'
(Nind, 2006: 188)

By documenting these challenges and reflections recommendations can be made on how the process of reviewing can be improved. Innovative strategies that reviewers have found useful could be developed further, and even subjected to evaluation where feasible. A second objective for this research is:

Research objective:

2. To assess the challenges reviewers have faced when doing systematic reviews of health promotion:

- How have these challenges been dealt with?
- With what success?

1.3 Context: defining health promotion

The term health promotion itself is sometimes used synonymously with health education. The relationship between the two can be complex, but generally health education can be viewed as a component of the much broader practice of health promotion (Tones and Tilford, 2001). The former comprises activities to provide information and advice either to individuals or groups of people in order for them to make healthy lifestyle choices, and is often measured in terms of its impact on attitudes, norms, and health-related behaviour. Health promotion, in contrast, encompasses a wider range of initiatives such as the development of health promoting legislation and policy (e.g. in schools, hospitals, the workplace). Its outputs, and therefore the measures by which it is judged successful, might include increasing the provision of healthy food choices in school canteens, or legislation to create smoke-free working environments. Ideally the two together should facilitate conditions in which people are able to exercise healthy informed choices. Their relationship has been summarised by Tones and Tilford (2001: 43) in the form of an equation: 'health promotion = health education x healthy public policy'.

Definitions and conceptions of health promotion have changed over the years as the discipline itself has evolved and broadened. Macdonald (1998) notes how in 1973 the then Canadian Minister of Health and Welfare made what appears to be one of the first references to the term in a report about new perspectives on health. The significance of this was that, for apparently the first time, it was explicitly acknowledged that the causes of ill-health could be attributed to non-medical origins, such as the environment and politics. The World Health Organisation over the years has increasingly focused on non-medical determinants of health. The Alma Ata Declaration (World Health Organisation, 1978), the Ottawa Charter (World Health Organisation, 1986), and the Jakarta Declaration (World Health Organisation, 1998), for example, all affirmed the importance of a broad perspective on the promotion of health including addressing socio-economic and political concerns. Generally there has been a shift away from the bio-medical model of 'disease prevention', to a more holistic approach in which health and well being are integrated in all aspects of a person's life.

Health promoters also talk less in terms of 'telling' people to change their behaviour (equated with what Tones and Tilford (2001) refer to as the 'preventive' model), and more in terms of empowering or encouraging them to make informed decisions about their health (the 'empowerment model'). For this to be achieved there has been recognition that health promotion needs to focus not only on the individual and their behaviour, but also to tackle the wider determinants of health, in order to engineer and sustain health promoting social, political and economic structures, as evident in this definition:

"Any combination of educational, organisational, economic and environmental support for conditions of living and behaviour of individuals, groups or communities conducive to health" (Green and Kreuter, 1991: 2)

Emphasis has therefore shifted from the relatively (and perhaps crudely titled) 'simplistic interventions' involving, for example, the provision of health education to individuals, to a broader profession which encompasses more complex activities aimed at, and actively involving, communities, regions or even countries via a range of different means (e.g. advocacy, lobbying, policy, legislation, mass media) (Campbell *et al*, 2000; Hawe *et al*, 2004; Pawson *et al*, 2005; Pawson, 2006b). This reflects UK Government health policy which is committed to tackling health inequalities and which urges Government departments to work collaboratively to ensure joined up policy (Acheson, 1998; Department of Health, 2004; Global Health Equity Group, 2009; Wanless, 2004) (for a review of health inequalities on the policy agenda in the UK see Kelly, 2006a). Thus, policies and strategies on the environment, transport, housing, and health, for example, should be integrated so that common aims can be met. For example, policies to reduce car use by increasing provision of public transport (particularly to rural areas) and creation of safer cycling facilities are likely to not only ease traffic congestion and benefit the environment, but will increase opportunities for people to participate in physical activity, thus likely reducing their risk of chronic disease.

1.3.1 Evaluating health promotion

The evolution of health promotion into a broader, more complex discipline presents challenges for its evaluation. As Tilford (2000) points out, measuring concepts such as empowerment, community participation and the development of healthy alliances is more problematic than measuring changes in health knowledge and health behaviours. Within health promotion, as in other areas such as social work research (Macdonald, 1997), and education (Hammersley, 2008; Oakley, 2003), there has been wide debate about the most appropriate methodologies to

evaluate effectiveness, particularly about the use of experimental methods, which, as explained earlier, occupy the higher echelons of the hierarchy of evidence.

As noted earlier there is strong, though by no means universal, support in health care for the RCT. Whilst it has been suggested that the RCT should be the 'gold standard' and used wherever possible (Loevinsohn, 1990; Macintyre and Petticrew, 2000; Oakley *et al*, 1995), others have commented that evaluating health promotion is a complex task and that RCTs, although advantageous, are not always practical or appropriate (Nutbeam, 1999; 2001; Speller *et al*, 1997; Weightman *et al*, 2005). Nutbeam (2001) suggests that complex multi-component interventions (e.g. directed at communities or regions using a range of media, delivered in a number of settings) are more likely to be effective in bringing about population health gains than 'single issue' initiatives (e.g. directed at individuals or small groups, using fewer media, delivered in a particular setting), but are much harder to evaluate.

The World Health Organisation (WHO) went as far as saying:

"The use of randomised control trials to evaluate health promotion initiatives is, in most cases, inappropriate, misleading and unnecessarily expensive"
(WHO European Working Group, 1998: 5)

However, although there has been some reluctance to use experimental methods in health promotion, the RCT has long been considered to be the optimal design for evaluation in related fields such as social policy and sociology, particularly in the US (Oakley, 1998; 2000; Oakley *et al*, 2003). Oakley (1998: 1239) discusses how a number RCTs were conducted in the US during the 1960s to evaluate the effectiveness of public policy:

"This history is conveniently overlooked by those who contend that randomised controlled trials have no place in evaluating social interventions. It shows clearly that prospective experimental studies with random allocation to generate one or more control groups is perfectly possible in social settings"

Support for the use of RCTs has not been limited to North America. There are examples of RCTs in Europe, including those which have evaluated the effectiveness of complex health promotion interventions. For example, the North Karelia Youth Programme (Vartiainen *et al*, 1991) was a large scale multi-component intervention in Finland evaluated using an RCT. It involved over 4000 participants, featuring a range of activities including classroom education, media campaigns, changes to nutritional content of school meals, health screening, and health

education initiatives in the workplace. Moreover, Bonell and Imrie (2001) cite a number of examples of RCTs which have been used to evaluate complex behavioural interventions to prevent HIV. The view that the RCT is impractical or inappropriate to test the success of health promotion is therefore not wholly tenable.

So far this chapter has set the context for this study by introducing systematic reviews and the concept of evidence-based health and systematic reviews. We have seen how considerable support and investment has been given to establish evidence-based health services, and have discussed some of the criticisms of this approach. To some extent these criticisms have been defended and it is beyond the scope of this investigation to try and resolve all of the debates. The long-standing debate about whether or not the RCT is appropriate to evaluate health promotion, for example, is likely to continue - albeit perhaps with less fervour than in previous years. This research intends to make a contribution by considering some of these issues specifically within the context of systematic reviewing. The next section, therefore, describes how and why systematic reviews appraise the methodological quality of evidence, and identifies unresolved issues for this research to investigate.

1.4 The process of quality assessment in systematic reviews

One of the ways in which systematic reviews attempt to adhere to rigorous methods is through identifying and accounting for biases in the methodology of included studies. The confidence that can be placed in the findings of an evaluation depends upon the evaluation design and the way it is conducted. A key stage in a systematic review, therefore, is to assess the quality of studies to ensure the recommendations and conclusions are based on sound evidence.

Before proceeding it is important to define quality. The term is often equated with validity, particularly internal validity, which can be defined as the degree to which a result of a measurement or study is likely to be true and free of bias (systematic errors) (Campbell and Stanley, 1966). In 2008 the Cochrane Collaboration chose to use ‘risk of bias’ as a replacement for the term quality, as it was considered to be a less subjective and more precise expression (Higgins and Green, 2008). This reflects only a recent change of policy and consequently in this thesis the term quality will be retained and used synonymously with internal validity.

A key distinction between internal and external validity is that the former is concerned with whether the observed effects are true for the people taking part in a study, whilst the latter is concerned about the extent to which effects observed in a study reflect what can be expected in the real world with different people, settings and times (i.e. its generalisability) (Campbell and

Stanley 1966; Cook and Campbell, 1979; Green and Glasgow, 2006). In this study the term external validity will be used separately to quality (internal validity).

1.4.1 Why is it necessary to assess quality?

At this point it is important to demonstrate empirically why it is necessary to assess quality. There are some striking examples of the impact of quality assessment on the results of systematic reviews (Egger *et al*, 2003). For example, two meta-analyses of the same trials of low molecular weight heparin (LMWH) compared to convention unfractionated heparin for the prevention of deep vein thrombosis (DVT) came to different conclusions (see Table 2).

Table 2 - Systematic reviews of the effectiveness of low molecular weight heparins (LMWH) vs standard heparin in the prevention of Deep Vein Thrombosis (DVT)

Author	Intervention	Quality assessment:	Conclusion
Leizorovicz *et al* (1992)	LMWH vs unfractionated heparin	not used	statistically significant benefit for LMWH
Nurmohamed *et al* (1992)	LMWH vs unfractionated heparin	used	no statistically significant benefit for LMWH

The review by Leizorovicz *et al* (1992) combined all of the relevant trials quantitatively, without examining their strengths and weaknesses, and concluded that there was significant benefit for patients taking LMWH. In the Nurmohamed *et al* (1992) review there was a significant reduction in the risk of DVT with LMWH, however when the analysis was restricted to trials which were judged to be methodologically superior, there was no statistical difference between the two drugs. This relates to the point made earlier, that poorer quality evaluations are more likely to over-estimate the effect of an intervention than higher quality ones. This underscores not only the importance of assessing quality, but also the need to investigate effective methodologies for doing so, as will be done in this study.

Given the importance of quality assessment in minimising bias it is of concern that not all reviews appraise quality, casting doubt on the extent to which they can be considered 'systematic'. A survey of 133 meta-analyses published in specialist and general medical journals between 1993 and 1997 found that only 41% had conducted quality assessment (Tallon *et al*, 2001). The picture was less optimistic in health promotion, where only around a third of the 400

reviews assessed in an audit reported any assessment of the methodological quality of the primary evaluations included (Peersman *et al*, 1999).

As more and more systematic reviews are published each year it is necessary to chart whether over time there has been an increase in the use of quality assessment. This will provide an indication of how trustworthy the findings are, allowing us to judge whether the evidence used to support policy and practice is sound. This can be assessed through a descriptive mapping of the methods used by systematic reviews of health promotion. More importantly, where quality has not been assessed it is crucial to ascertain whether there are any barriers, and if so, how these might be removed. Systematic reviewers themselves should be able to elucidate these issues. This will be of practical value to the field as recommendations can be made for more effective systematic reviewing. Therefore an objective for this research is:

Research objective:

3. To assess the extent to which systematic reviews of health promotion tend to assess the quality of included studies:

- What are the barriers to, and facilitators of, quality assessment?

1.4.2 Is quality assessed systematically?

Peersman *et al's* audit of health promotion reviews found that only a third of reviews explicitly reported the quality assessment criteria used, suggesting that systematic reviews can be ambiguous about the basis upon which they judge studies. For example, in a systematic review of the effectiveness of interventions to prevent teenage pregnancy, Kirby *et al* (1994: 348) did not report a formal quality assessment process but did critique the evidence. Comments ranged from general appraisal of the studies:

"The evaluation was very rigorous: it had random assignment, large sample sizes, high consent rates, short and long term follow-up, low drop out rates, and appropriate statistical analyses"

to specific criticisms:

"sample sizes for some subgroups were too small for reasonable power"

Their summary of the evidence base was that the evaluation of pregnancy prevention interventions has serious limitations prohibiting definitive conclusions about effectiveness. Whilst a critical stance on the evidence is one of the merits of this review, absence of explicit details about quality assessment procedures is problematic. If studies are not appraised consistently there is the danger that some are singled out for criticism over others. The reader is unable to judge whether or not a systematic approach was followed and whether the findings of the review are potentially biased as a consequence. It is also at odds with recommendations from guidelines and key texts on the conduct of systematic reviews (Egger *et al*, 2001; Centre for Reviews and Dissemination, 2009; Higgins and Green, 2008). It is important to assess the extent to which this occurs, through descriptively mapping the methods used by systematic reviews of health promotion. If it is found to be a common phenomenon then relevant organisations can be encouraged to strengthen their recommendations to systematic reviewers. Therefore an objective for this research is:

Research objective:

4. To assess the extent to which quality assessment is conducted and reported in a 'systematic' manner.

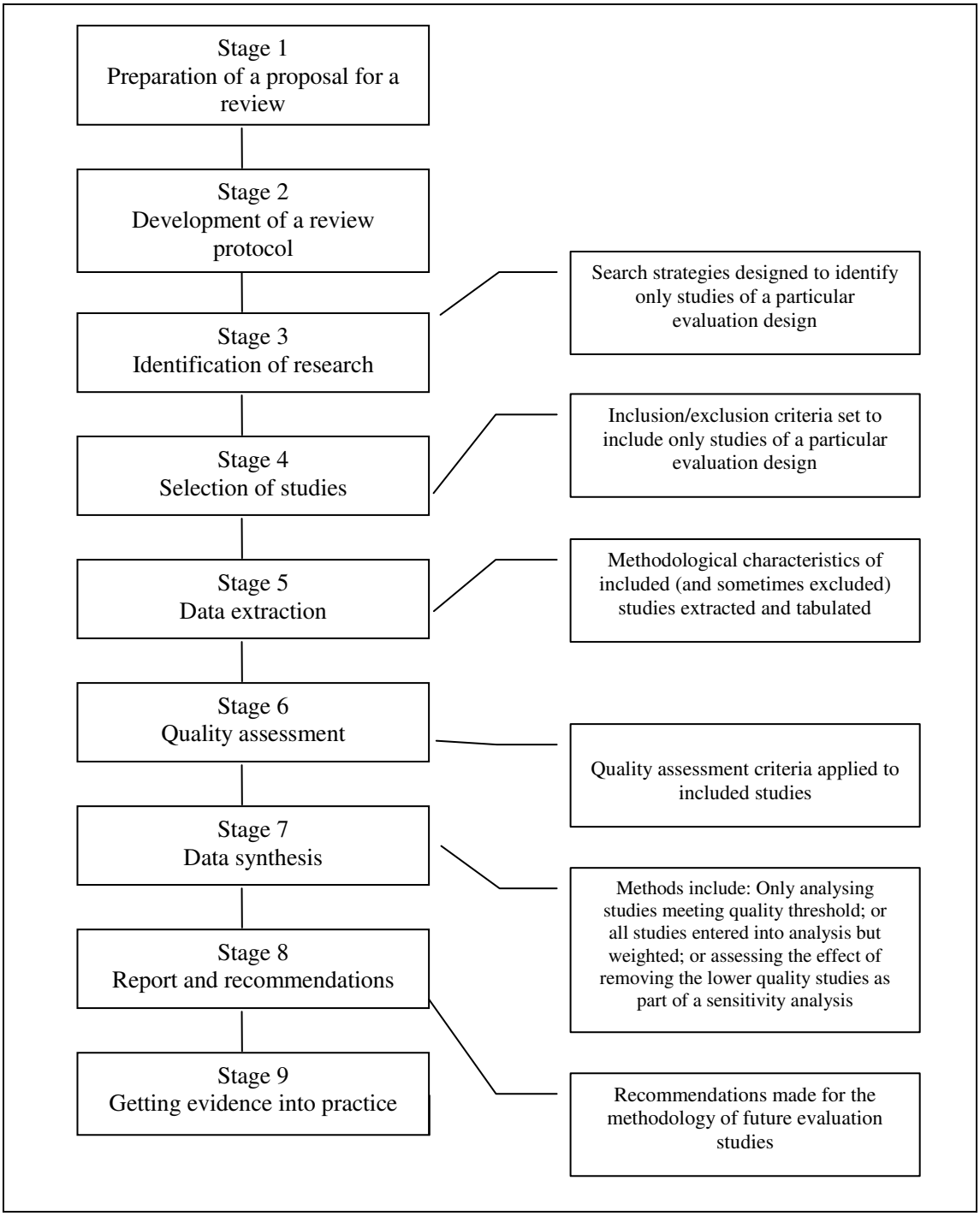
- Do systematic reviews of health promotion tend to apply the same set of criteria to each study?
- Do systematic reviews of health promotion tend to single some studies out for criticism over others?
- Do systematic reviews tend to criticise studies for specific methodological flaws without having formally appraised them?

1.4.3 How is quality assessment used?

Figure 2 shows how considerations of methodological quality can take place at key stages in the review process. Quality might be considered at one or more of these stages, such as at the start of the review when locating the evidence and screening studies for inclusion (Stages 3 and 4); and/or as a formal appraisal exercise involving the application of *a priori* criteria (Stage 6); and/or when synthesising the results of the studies (Stage 7).

For example, from the outset it may be decided that only studies of a particular design will be relevant to a review. Accordingly, studies of this kind will be prioritised during literature searches, and when sifting the results of the search only those designs will be included.

Figure 2 - Considerations of study quality at different stages of a systematic review



Adapted from Centre for Reviews and Dissemination (2001)

Given the impact that quality assessment can have on the findings of a systematic review, it would be reasonable to expect reviews to consider quality at as many stages as possible. Whilst it may not always be appropriate to restrict inclusion to studies of a particular design, guidelines and key texts recommend that it is important to formally assess the quality of the included studies (i.e. Stage 6), and to then take this into account when analysing the results and forming

conclusions (i.e. Stage 7 and 8) (Detsky *et al*, 1992; Egger *et al* 2001; Higgins and Green, 2008; Jüni *et al*, 2001; Moher *et al*, 1998). As will be discussed in the next section, there is potential for over-estimation of the intervention effect if the results of poorer quality studies are accepted uncritically. The benefit of critically appraising the studies included in systematic reviews is therefore maximised when the results reflect the strengths and weaknesses of the included studies.

There are at least three ways of integrating quality judgement in the analysis of results of a systematic review (Detsky *et al*, 1992; Egger *et al*, 2001).

1. Quality weighting

'Quality weighting', as the title suggests, allocates more weight to studies of higher methodological quality enabling the poorer studies to exert less influence on the results. For example, studies entered into a meta-analysis can be weighted according their quality score, or some other classification of their quality (although quality scores, at one time commonly used, have become generally discouraged - Jüni *et al*, 1999; 2001; Kunz and Oxman, 1998).

2. Threshold approach

A 'threshold approach', in contrast, prioritises a subset of studies deemed to be of sufficient quality to support the results of a review. There are different ways to designate such a threshold. For example, if a scale has been used to assess quality then only studies scoring above a pre-specified threshold may be considered for analysis. Similarly, if a checklist has been used then only studies meeting certain pre-defined criteria may be included.

3. Sensitivity analysis

A 'sensitivity analysis' explores the effects of the addition/removal of poorer quality studies on the results and conclusions of a review.

Reviews which conduct a formal assessment of quality and then integrate their judgement into their findings will arguably be more credible than those which do not. However, few studies have charted the extent to which the results of systematic reviews reflect the quality of the evidence (Moja *et al*, 2005), particularly in health promotion. During the agenda-setting interviews for this study it was commented that the various methods (discussed above) of integrating quality judgement into the results of a systematic review have advantages and

disadvantages. For example, only analysing studies meeting a pre-defined threshold of quality may reduce the workload for the systematic reviewer, but studies at the excluded margin of the threshold may not be drastically different in terms of quality than those on the margins of the include threshold. Had they been included the results and conclusions of the review may potentially be quite different. It is not currently clear which of the various methods tend to be used by systematic reviews, or even whether reviews combine more than one type of approach. A mapping of the methods used by systematic reviews of health promotion is necessary to answer these questions. The mapping would shed light on whether guidelines on the conduct of systematic reviews are being observed (Centre for Reviews and Dissemination, 2009; Higgins and Green, 2008). Therefore, an objective for this research is:

Research objective:

5. To assess how systematic reviews of health promotion use quality judgements:
- Do the findings and conclusions of systematic reviews reflect the strengths and weaknesses of the included studies?
- If so, by which methods? (e.g. quality thresholds; quality weighting, etc)
- Is there consensus on the most appropriate method?

Having justified the rationale for assessing quality in systematic reviews, and identified the need to investigate the extent to which this is routine, the next section discusses a number of outstanding issues regarding the process of assessing quality.

1.5 Which quality assessment criteria are used by systematic reviews of health promotion?

1.5.1 Empirically demonstrated threats to internal validity

Much empirical research has been conducted to identify significant sources of bias in clinical trials. Inadequate procedures in three areas have been demonstrated to bias the results of clinical trials:

1. Inadequate randomisation of participants to study groups, and failure to conceal allocation ('selection bias').

Some investigators assign participants to groups using methods that are not considered 'true' randomisation, such as using alternate numbers (Kjaergard *et al*, 1999). This is unlikely to achieve the even distribution of participants to study groups necessary to prevent selection bias.

Furthermore, if randomisation is not adequately concealed then subversion can happen, particularly if an investigator favours a particular individual to receive the experimental intervention, as has been the case in clinical trials (Schulz, 1995). Inadequate concealment can inflate the effectiveness of an intervention by up to 30% (Jüni *et al,* 2001; Schulz *et al*, 1995). Reviews which include such studies may erroneously conclude that an intervention is more effective than it really is, potentially misleading policy makers and practitioners into implementing initiatives of little demonstrated effectiveness.

2. Insufficient blinding of participants to their study group ('performance bias')

Knowledge of group assignment can influence the intervention recipient's perceptions about the degree to which they are benefiting from it. This is a particular problem where the outcomes are self-reported and therefore less objective (Wood *et al*, 2008). Recipients may over- or under-report changes in symptoms, attitudes, beliefs or behaviour depending upon whether they know they are receiving the experimental intervention, or the comparator. It has been demonstrated that effects can be over-estimated by up to 9% in RCTs with lack of blinding (Pildal *et al*, 2007). Insufficient blinding of people who assess outcomes of an intervention (e.g. researchers, clinicians, laboratory staff, etc) to the study group assignment can also influence the data recorded, particularly if an element of subjective judgement is required (Boutron *et al*, 2007; Flay, 1986). This is referred to as detection bias.

3. Procedures for dealing with withdrawals from the study ('attrition bias')

Significant numbers of withdrawals from a trial can bias results, particularly if those who leave are systematically different in characteristics from those who remain. The problem is exacerbated if there is disproportionate drop-out between the intervention and comparator groups. Empirical studies of the effect of missing data resulting from attrition have had mixed results (Kjaergard *et al*, 2001; Schulz *et al*, 1995; Tierney, 2005). Guidelines on systematic reviewing nonetheless recommend this to be a key consideration in the assessment of methodological quality on theoretical grounds (Higgins and Green, 2008; Centre for Reviews and Dissemination, 2009).

Given the potential of these attributes to bias results it is crucial that they are taken into account in any assessment of quality. An important question is whether, in practice, they are. There is some evidence that this is the case. The results of a review of 25 scales and 9 checklists used to assess the quality of RCTs in health care found that, with only one exception, all of the instruments included items based on "accepted criteria" (defined as being based on textbooks on

clinical trials) (Moher *et al*, 1995; Moher *et al*, 1999). However, there was great variability in the weight given to the three methodological attributes thought to be most strongly related to bias. For example, one scale contained only these three items and assigned equal weight to each. In contrast, another scale contained a total of 29 items and gave just 2.5% weight each to both randomisation and blinding, and 3.1% to withdrawals. One scale didn't even assess any of the three attributes at all.

Another investigation sought to identify the characteristics of instruments used to assess the quality of non-randomised studies (Deeks *et al*, 2003). A total of 194 tools were identified from a systematic search of electronic databases, reference lists, and contact with experts in the field. In contrast to the findings of Moher *et al*'s study, over two thirds did not describe how the criteria were chosen, and whether they covered attributes empirically associated with bias. The majority of the remaining third chose criteria according to the recommendations from methodological literature, with only a minority making use of an expert panel or survey.

What is clear from these two studies is that, in practice, there is variability in the quality assessment criteria employed in systematic reviews of health care interventions, and the criteria do not always take into account key threats to internal validity. It is important to extend the investigations cited above to health promotion, as there has been little published research in this area. The importance of this issue was endorsed by two of the interviewees in the agenda-setting exercise undertaken in this study. They queried whether or not systematic reviews routinely provide a rationale for the quality criteria they employ, and where a justification is given whether it is based on empirical evidence of bias.

Although it is recommended that any assessment of quality of clinical RCTs should include examination of the three key methodological attributes associated with bias, the question arises 'Are such criteria applicable to evaluations of health promotion interventions?' The next section discusses this in greater detail.

1.5.2 Applicability of quality assessment criteria to health promotion?

As discussed earlier, it has been argued that there are challenges in evaluating health promotion interventions, particularly using experimental methods. However, it would be misleading to assume that such methods are wholly impractical and rarely used. Health promotion interventions differ in a number of ways from health care interventions, prompting the question: Which dimensions of internal validity are particularly relevant in the assessment of quality of health promotion evaluation?

The answer to this question depends on the type of health promotion intervention evaluated. Some interventions (such as the introduction of new policy and legislation) are rarely amenable to the inclusion of a control or comparison group (randomly allocated or otherwise). For example, in the UK it would be difficult to identify a control group to assess the impact of lowering the age of consent for male homosexual intercourse on HIV prevention (Bonell and Imrie, 2001). In this situation an evaluator might instead use a case study design, perhaps measuring the impact of the introduction of the new policy in a particular area or region before and after its implementation.

Where a control or comparison group can be included, random allocation (concealed or otherwise) of participants to groups may pose problems. Baranowski *et al* (1990) evaluated an intervention to promote physical activity among Black-American families in the US, randomly allocating them to either an intervention in which they attended fitness sessions at a local community centre or a control group. Attendance at the centre was poor and was partly explained by the fact that some of the families randomised to the intervention would have been more motivated to attend if they could have gone with their neighbours, who unlike them, had been randomised to the control group. Given that the main benefit of randomisation is the even distribution of factors known (and unknown) to be related to the effectiveness of an intervention (Altman and Bland, 1999; Rossi *et al*, 1999; Stephenson and Imrie, 1998), in its absence the issue of comparability between study groups at the start of the evaluation is paramount in order to assess the risk of selection bias. Therefore, for non-randomised trials the comparability of groups at the start of the study (i.e. at 'baseline') would be an important marker of quality. Whether this is assessed by systematic reviews of health promotion is an issue for this investigation.

Blinding of participants to their group assignment in controlled studies as an indicator of validity, although important, is not likely to always be achievable in health promotion (Flay, 1986). For example, it would be harder to mask experimental classroom health education curricula than it would be in the evaluation of a drug where patients in the intervention and control groups are given identical looking tablets but where only the former receive the tablet containing the active ingredient. It is necessary to investigate whether blinding is always considered to be an inappropriate criterion to judge the quality of health promotion, or whether there are circumstances when it is applicable.

Attrition bias poses a particular problem for health promotion. This is particularly the case for interventions which occur in community settings where study participants are often highly transient and consequently more difficult to trace, particularly over long periods of time (e.g.

commercial sex workers, or young people from disadvantaged socio-economic backgrounds) (Coyle *et al*, 2006; Gwadz and Rotheram-Borus, 1992). The extent to which attrition poses a threat to the internal validity of such studies is therefore an important consideration in quality assessment of health promotion interventions. It is important to investigate whether systematic reviews of health promotion assess attrition bias.

A further problem (specific to controlled trials) is the potential for the control/comparison group to become exposed to the intervention, particularly in settings where participants in the intervention and control groups are highly likely to interact with each other (e.g. within schools), thus confounding the results (Torgerson, 2001; Keogh-Brown *et al*, 2007). A potential solution would be to allocate participants to study groups in clusters (e.g. a group of general practice surgeries; a block of housing units). However, cluster trials have their own idiosyncrasies. For example, there would need to be enough clusters in the sample to enable a statistically significant intervention effect to be detected, in which case use of a statistical power calculation to determine the appropriate sample size would be an appropriate marker of quality (Elbourne and Campbell, 2001). An appropriate method of data analysis would be required to ensure that the unit of analysis matches the unit of allocation to study groups. That is, if schools are allocated to intervention and control groups then schools, rather than individual pupils, must be the unit of analysis (Killip *et al*, 2004; White and Thomas, 2005).

Other threats to the validity of health promotion evaluations (whether experimental or not) include poorly constructed data collection instruments (e.g. questionnaires; interview schedules) which have not been adequately validated (Oppenheim, 1992; Bailey, 1994), or inappropriate methods of data analysis (Miles and Huberman, 1994; Rose and Sullivan, 1996). Another issue is whether the length of the evaluation is sufficient enough to measure changes in all of the relevant outcomes. For example, given that behaviour change takes time to adopt and maintain (Prochaska *et al*. 1994) it is important that a suitable enough follow-up period is adopted, otherwise it might be wrongly concluded that an intervention failed to have an effect when, in actuality, an effect would have been observed had the evaluation lasted longer. All of these issues pose significant potential problems and ideally should be taken into consideration in systematic reviews. Whether or not they are is currently unclear, and is an issue for this study to investigate.

What emerges from this discussion is an appreciation that health promotion may be evaluated by a diverse range of study designs, and that the 'key' criteria to assess the quality of health care evaluations may not always applicable to health promotion. Methodological attributes of controlled trials such as concealed randomisation and blinding, although important criteria for

the assessment of quality of health care evaluations, may not always be possible. Whilst some of these issues are not unique to health promotion they are nevertheless common problems faced by evaluators in this area, and present considerable challenges to assessing quality of studies in systematic reviews.

The issues discussed here have received little attention in the literature, and there are unanswered questions. Which criteria have systematic reviewers in health promotion tended to use? What justification do they provide for their choice of criteria? Is there any consensus on the most appropriate quality criteria to apply in systematic reviews of health promotion? The use of health promotion as a case study is all the more necessary given the paucity of methodological work in this area. A first step would be to assess which criteria have been employed in published systematic reviews of health promotion, through methodological mapping. This would illustrate consensus in terms of criteria that have been actually used. A second step would be to ask systematic reviewers to reflect on their choice of criteria, and to discuss whether they think there is consensus. Therefore, objectives for this research are:

Research objectives:

6. To assess the criteria that systematic reviews of health promotion use to assess the quality of included evidence:

- Which criteria are used?
- Why have these criteria been chosen?
- Do these criteria address acknowledged threats to internal validity?

7. To assess whether there is consensus on the criteria by which health promotion evaluations should be assessed in systematic reviews.

1.6 The assessment of external validity in systematic reviews of health promotion

As mentioned earlier, external validity is concerned with the degree to which the results of an observation hold true in other settings. The literature on external validity in scientific research extends back to the 1960s (Campbell and Stanley 1966; Campbell, 1986). More recently, texts on the theory and practice of health care and health promotion have also discussed its importance (Green and Glasgow, 2006; Green and Kreuter, 1999; Tones and Tilford, 2001). It could be argued that the assessment of external validity is particularly important in health promotion as interventions are often complex, multi-faceted and context specific. Yet

commentators have noted the comparative neglect of external validity in health research (Flay, 1986). There has been a disproportionate focus on evaluating 'efficacy' in optimal circumstances, rather than 'effectiveness' in the real world (Green and Glasgow, 2006).

In recent years, however, guidelines for the conduct of systematic reviews have emphasised the importance of considering external validity, in terms of the appropriateness and generalisability of interventions (Armstrong *et al*, 2008; Jackson *et al*, 2004; Jackson and Waters, 2005; Centre for Reviews and Dissemination, 2009; Petticrew and Roberts, 2006). The aim of assessing external validity in a systematic review is to make it more meaningful and relevant to stakeholders. Whether reviews always achieve this, however, is questionable as the literature on systematic review methodology has tended to prioritise issues relating to internal validity (Speller *et al*, 1997). Consequently there is a gap where methodological research is needed to guide systematic reviewers. The question arises 'What aspects of external validity should systematic reviews assess?' Three come to mind.

1.6.1 Generalisability

The first is generalisability, that is, the extent to which study findings are applicable to local populations and settings (Green and Glasgow, 2006). A practitioner, for example, can assess not only whether a programme or intervention is effective, but whether it is likely to be effective in their own locality. To help them make this judgement they may need to know the characteristics of the study population (e.g. age, gender, socio-economic status, ethnicity etc), and the setting (e.g. school, workplace, health care setting). For systematic reviews to be of maximal use to stakeholders it is essential for them to discuss generalisability (Bonell *et al*, 2006). However, it is not clear whether they routinely do, suggesting an issue to be investigated in this study.

1.6.2 Replicability

Replicability is the degree to which an effective intervention might be implemented by others outside the context of an evaluation (Flay, 1986). If the practitioner is confident that the results are generalisable to their population and setting they may plan to implement the intervention locally. There has been considerable interest in the replication of effective health promotion interventions, particularly in the US where the National Institute of Health has funded a number of replication studies of HIV prevention interventions (Bell *et al*, 2008). Systematic reviews can help identify which interventions are effective and therefore eligible for replication. Ideally they should provide relevant information to allow decisions to be made on which aspects might need adapting to suit the local context, such as cultural relevance. It is not clear, however,

whether they commonly do, or what the key markers of replicability are. Again, this is an issue to be investigated in this study.

1.6.3 Quality of the intervention

Thirdly, the quality of the intervention itself can be indicative of the external validity of a study. Speller *et al* (1997) notes how systematic reviews often devote more attention to the quality of the evaluation than to the intervention. Consequently they make recommendations based on high quality evidence, but for the effectiveness of potentially dubious interventions.
The importance of this issue is underlined by the fact that interventions which are not implemented as designed, or which are inappropriate for a given health issue can influence the results of an evaluation (Herbert and Bø, 2005). It has been therefore been recommended that systematic reviews should routinely assess the quality of the intervention (Herbert and Bø, 2005). It is unclear, however, whether this recommendation has been adopted, necessitating investigation.

1.6.3.1 Potential markers of intervention quality

A literature on quality assurance in health promotion has emerged over recent years, but with little apparent consensus on markers of quality (Ader *et al*, 2001; Catford, 1993; Evans *et al*, 1994; Speller *et al* 1997, Tones 2000; van Driel and Keijsers, 1997). What might be the potential indicators of a high quality health promotion intervention? Figure 3 lists some potential candidates.

Figure 3 – Potential markers of intervention quality

- Was the intervention based on a needs assessment?
- Was the intervention designed with input of its target population?
- Was the intervention devised according to theoretical principles?
- Was the intervention designed according to the principles of health promotion?
- Was the intervention implemented as planned?
- Does the intervention comply with ethical principles?

A high quality intervention might be, amongst other things, one that is adequately resourced, ethical, policy-relevant, appropriate to the principles of health promotion, and meaningful to its recipients. It might also be one that is based on a plausible theory of change. A great deal has been written about the contribution of theory to health promotion (Bonell and Imrie, 2001; Fisher and Fisher, 2000; Green, 2000; Rothman *et al*, 2004; Tones and Tilford *et al*, 2001;

Turner and Shepherd, 1999; Wight *et al*, 1998). The aim of using theory is to predict and explain the important social, psychological, cultural and environmental determinants of health and health behaviours. It might be used to predict how people may respond to an intervention, guide the selection of indicators to demonstrate effectiveness, and may help us to interpret the findings of evaluative studies. The argument in favour of using theory is that the success or failure of an intervention is easier to explain if based upon a relevant concept of health related change. Without it the intervention may resemble a 'black box', with no understanding of its guiding principles. It is not clear, however, whether systematic reviews in health promotion assess the extent to which the included interventions are theory-based, and whether the reviews themselves use theory to attempt to explain their findings.

Another marker of quality might be whether or not the intervention observes the key principles of health promotion as set out in the Ottawa Charter (World Health Organisation, 1986) and the Jakarta Declaration (World Health Organisation, 1998). Namely, it should be equitable, participative, collaborative, and one which empowers people, communities and organisations to make health promoting changes (Tones and Tilford, 2001). It could be argued that interventions which coerce or manipulate people into changing aspects of their lives instead of using empowerment approaches are undesirable and therefore regarded to be of poor quality. An example might be interventions which use persuasive arguments to encourage abstinence from premarital sex among young people. Evaluation of this approach has found it to be ineffective and even harmful (Oakley *et al*, 1995).

In determining what might constitute a good quality health promotion intervention one might look to the principles of quality under-pinning health care. For example, Mullen *et al* (1985), in an early example of a meta-analysis in health, judged the quality of patient education interventions for chronic disease management according to adherence to educational theory, including individualisation, feedback and reinforcement. Since education is a component of many health promotion interventions these might be appropriate markers of quality.

Another pertinent question is whether poor quality interventions, particularly ethically dubious ones, should be included in systematic reviews at all? Some would argue that all evidence be included in reviews to minimise publication bias. Omission of particular studies may drastically alter the overall estimate of effectiveness, resulting in a false representation of the extant evidence in a given area. Others might argue a pragmatic approach whereby all studies are included but with suitable caveats in the discussion and conclusions of review to alert users that certain interventions are ethically questionable. This assumes that ethics is a homogenous concept, when in fact it might comprise a number of distinct issues.

The degree to which an intervention is delivered as intended (sometimes referred to as fidelity) is another aspect of intervention quality (Rychetnik *et al*, 2002). If not implemented according to design the intervention may be ineffective, or even harmful (Dane and Schneider, 1998; Dumas *et al*, 2001). Lack of monitoring and audit data to confirm whether an initiative was delivered according to its protocol makes it difficult to fully explain the outcomes. For example, if an intervention was less successful than expected was it because it was under-resourced, or was it because it was poorly designed? This becomes more of a problem when the intervention is complex, involving a number of different providers and settings (Herbert and Bø, 2005), as is often the case in health promotion.

1.6.4 Summary

In summary, generalisability and replicability are considered important aspects of external validity in the literature but it is not clear whether systematic reviews of health promotion routinely assess them. There is also a need to establish which markers of generalisability and replicability are commonly assessed. This study also goes beyond generalisability and replicability to incorporate issues concerning the quality of the intervention itself. There are a number of potential markers of intervention quality, but little work appears to have been conducted to establish which have been considered important by systematic reviews. It is therefore necessary to map the markers of which have been used in reviews, and to discuss their importance and relevance to health promotion. The outcome would be recommendations for the production of potentially more meaningful and useful systematic reviews. Therefore, an objective for this research is:

Research objective:

8. To assess the extent to which systematic reviews of health promotion assess the external validity of included studies:

- For what purpose do systematic reviews of health promotion assess external validity?
- What are the key markers of external validity?

1.7 Who is involved in the production of systematic reviews of health promotion?

1.7.1 Which stakeholders participate in systematic reviews?

An issue that has received comparatively little investigation is the type of people who tend to be involved in the production of systematic reviews, in general and in health promotion. A

reasonable, though simplistic, assumption would be that systematic reviewing is an academic endeavour. Systematic reviews are a form of research, and research is predominantly done by academics. This assumption is partly based on my personal experience. I work as part of a team of researchers in a University department which holds a contract to routinely produce systematic reviews for NICE on the clinical-effectiveness and cost-effectiveness of various health care treatments. We collaborate with similar teams in other universities, including the EPPI-Centre in London, and the CRD at the University of York, all of whom produce systematic reviews directly for policy making bodies. In my experience, then, the evidence needs of policy makers are being met to a large extent by researchers.

However, this does not necessarily preclude the fact that other stakeholders may be involved in the production of systematic reviews. In recent times there has been a move towards a more inclusive approach to health services research. The Government's health research strategy 'Best Research for Best Health' (Department of Health, 2006) makes a commitment to supporting the production of systematic reviews, and involving health professionals in all aspects of research. The strategy also stresses the importance of involving the public in research. This reflects a long-standing commitment to increase public participation in health research and decision making (Goodare and Smith, 1995; Goodare, 1999; Oliver *et al*, 2004; Royle and Oliver, 2004). The rationale for public participation in health research has been made clear: they can help ensure the research question, the intervention and its outcome measures are as relevant and meaningful as possible to those for whom the research is intended to serve. Some would even say that their involvement is an ethical imperative (Harden and Oliver, 2001). For example, the NIHR HTA Programme provides the public with the opportunity to identify and prioritise research topics (Oliver *et al*, 2004). Similarly, NICE operates the Patient and Public Involvement Programme (PPIP) which advises on lay and community input to its public health work (National Institute for Health and Clinical Excellence, 2009). The Cochrane Collaboration, through the Cochrane Consumer Network, encourages consumers to comment on the appropriateness of review protocols and completed systematic reviews (Cochrane Consumer Network, 2008). Whilst in principle there is commitment to including a range of stakeholders in the production of systematic reviews, it is far from clear whether this is a reality. This will therefore be assessed by this study.

1.7.2 Who might be best placed to assess quality in a systematic review?

There is also little discussion in the literature of the pre-requisites of an effective systematic reviewer, in terms of background, professional or academic status, level of skills, knowledge and experience. For example, what background might be most appropriate for specific tasks of a

systematic review, such as quality assessment? This is arguably one of the most demanding tasks of a systematic review as it requires an understanding of evaluation methodology and principles of validity and reliability. There are few, if any, published accounts of practitioners or the public participating in tasks such as data extraction, quality assessment, and synthesis of data. In a survey of 20 health promotion managers in the UK many reported that, in their attempt to follow an evidence-based approach to practice, they found it difficult to assess the validity of the research they accessed (Learmonth and Watson, 1999). This suggests that, unless adequate training is given, practitioners might not necessarily be in the best position to conduct some of the key tasks of a systematic review, an issue that warrants further investigation (training is discussed further in Section 1.8.2).

Hammersley (2001: 548) suggests that quality assessment is not a procedure that can be performed without specific knowledge of the particular topic or issue being investigated:

"Assessing the likely validity of the findings of a study never simply amounts to assessing its research design. One does not have to believe that validity is a matter of insight, intuition or standpoint to doubt the value of the procedural approach to assessing it which underlies systematic review…using fixed, standard criteria specifying a hierarchy of research designs ignores these sources of variation. It neglects the extent to which assessing the validity of studies' findings is a matter of contextually sensitive judgement...the assumption is that studies can be assessed in purely procedural terms, rather than on the basis of judgements which necessarily rely on broader, and often tacit, knowledge of a whole range of methodological and substantive matters"

Hammersley's position is not unlike that of Ray Pawson (Pawson, 2006a; Pawson *et al*, 2005) who views the assessment of quality within a realist synthesis as being highly contextual. Quality is assessed in terms of relevance to the theory of the intervention, and rigour with which a study makes a credible contribution to the test of that theory. It is suggested that both of these criteria can only be judged at the point of the synthesis. That is, once all of the evidence has been read, assimilated and analysed (as opposed to a systematic review in which quality assessment takes place outside of the context of any assimilation of the nature of the evidence). Implicitly, due to the complex nature of realist synthesis, this is only likely to be possible for someone already highly familiar with the topic from the outset (indeed, some could say that even without initial familiarity with the topic, the process of building a realist synthesis would render them an expert them by the stage at which quality is considered).

Returning to Hammersley, it is implicit in his texts exactly how substantive knowledge would improve the process of quality assessment. Furthermore, he does not speculate on what the difference to the conclusions of a systematic review might be with the presence or absence of substantive knowledge. He seems to argue against the need for objectivity and overlooks the possibility that a strong connection with a topic area may prevent a reviewer from being impartial when assessing the quality of studies. A potential disadvantage of using topic experts is that they 'may have pre-formed opinions that can influence their assessments' (Higgins and Green, 2008), a concern that also emerged in the agenda-setting interviews conducted in this study. Although declaration of conflict of interests is now a common procedure in the production of systematic reviews the concern is that, even where no major conflicts are declared, their judgement may be subtly compromised. They may end up reviewing their own studies (which, as explained in the Introduction to this thesis, has been my experience), or those of their colleagues. The results of their review may be contra to their own personal/professional view, which may consciously or sub-consciously influence the quality judgement they make.

Whilst many systematic reviewers in my position are trained in the specialist skills required to conduct reviews, due to the changing needs of policy makers they may have little or no substantive knowledge of the varying topics they are requested to review. Conversely, experts in given topic area may lack the skills to carry out systematic reviews of topics they know best. An obvious solution would be for systematic reviewers and topic specialists to work together to combine both methodological expertise and substantive knowledge, thus assuaging the concerns raised above about objectivity. Collaborative team working is increasingly a condition of funding for systematic reviews. Funders such as the Medical Research Council and the NIHR HTA Programme strongly encourage multi-disciplinary collaboration to those applying for funding for secondary research (NIHR Health Technology Assessment Programme, 2009). Yet little has been documented about the advantages and disadvantages of team working, and the feasibility of convening and running such teams within the resources and timescales that systematic reviews in health promotion are commonly allocated. This issue will therefore be explored in this study.

1.7.3 Summary and research agenda

What emerges from this discussion is the need for investigation into a number of issues relating to the characteristics of the people who conduct systematic reviews of health promotion. Few studies have documented who becomes involved in systematic reviewing, for what reasons, and what challenges and successes they faced. To what extent are systematic reviews produced by people with expert knowledge of the topic area, by those whose specialist skills are in

systematic reviewing, and by the two together? These issues will be investigated first via a mapping of published systematic reviews of health promotion. The map will chart the characteristics of the authors in terms of background and professional status. Secondly, the views and experiences of systematic reviewers themselves will be sought. This research will clarify whether the ideology of inclusiveness in evidence-based health is reality or rhetoric, and make recommendations, where necessary, about strategies for involving a variety of stakeholders and for effective team working. An objective for this research is therefore:

Research objective:

9. To assess which types of people commonly participate in the production of systematic reviews of health promotion:

- Who does reviews (e.g. academics, health and other professionals, lay people), and what is their rationale for doing them?
- Who performs quality assessment in systematic reviews? (e.g. people who specialise in producing systematic reviews; people who specialise in the topic area being reviewed; combinations of these)
- To what extent are systematic reviews the product of collaborative teams? What are the advantages and disadvantages of collaborative team working?

1.8 How do reviewers learn to do systematic reviews of health promotion?

1.8.1 What learning opportunities exist for learning to do systematic reviews in health promotion?

There currently appear to be a number of options available to those wishing to become skilled in systematic reviewing. Introductory training courses are offered worldwide by the Cochrane Collaboration, and in UK both the CRD and the EPPI-Centre run entry-level courses. The EPPI-Centre's course is part of the Economic and Social Research Council (ESRC) National Centre for Research Methodology programme which was set up to increase capacity in different forms of research, including evidence synthesis (this term being an increasingly used synonym for systematic review) (Wiles and Bardsley, 2008). As mentioned earlier, written learning resources are available including a detailed guide by the Cochrane Collaboration (Higgins and Green, 2008). The Cochrane guide incorporates specialist guidelines for health promotion produced by an international taskforce (to which I contributed a section on quality assessment) (Armstrong *et al*, 2008; Jackson *et al*, 2004; Jackson and Waters, 2005).

The principles and practice of evidence based-health are also taught as part of the curriculum in nurse and medical training (General Medical Council, 2003; Parkes *et al*, 2001), and in post-graduate qualifications in health services research and public health (London School of Hygiene and Tropical Medicine, 2008). It is would seem then, on the surface, that people wanting to learn systematic review skills can take advantage of a variety of resources. However, one of the interviewees who participated in the agenda-setting exercise in this study suggested that it is only relatively recently that training opportunities on the production of systematic reviews have increased, particularly in the social sciences. Furthermore, it was commented that text books on research methodology have traditionally lacked guidance on how to critically assess evaluation designs. These comments raise important questions about whether the provision of training and education for systematic reviewing in health promotion, and in more broadly in the social sciences, is adequate, something that has been questioned in the literature (Oakley *et al*, 2005). Are there currently any significant barriers to accessing support and training? What methods have reviewers found to be most useful and acceptable? In the absence of training and education opportunities how have systematic reviewers learned the process? This study will attempt to answer these questions through asking systematic reviewers to discuss their experiences of 'learning the ropes'.

1.8.2 How do systematic reviewers learn critical appraisal skills?

There is a dearth of academic literature about the process of learning systematic review skills. In searching the literature I found very few publications which discuss and reflect on different learning strategies, particularly critical appraisal (a term that is often used in the field synonymously with quality assessment). The Critical Appraisal Skills Programme (CASP), initiated in the early 1990s to help professionals and health care consumers learn to appraise evidence (Critical Appraisal Skills Programme, 2008), was evaluated by Milne and Oliver (1996). They found that the workshop participants, comprising health care consumers, rated the workshops very highly in terms of learning and satisfaction. It was concluded that these brief introductory workshops are feasible to run, but acknowledged that they are limited in the extent to which participants can become experts.

Oliver and Peersman (2001) evaluated a series of brief critical appraisal workshops for health promotion practitioners and purchasers. A thoughtful account is given of the needs and opinions of practitioners at that time (mid 1990s), and it was concluded that health promotion practitioners can develop basic appraisal skills, but need a supportive environment to apply them in their work. Whilst informative, these two small-scale studies are likely to have been overtaken by changes in practice and advancements in methodology mentioned earlier in this

chapter. Arguably the issues facing systematic reviewers today are more challenging, calling into question the extent to which practitioners and health care consumers, likely to be at a disadvantage to those with an academic background, can effectively address them (as discussed in the previous section).

West *et al* (2002: 79) discuss how specialist training might be required to handle the complexity of certain evaluation designs:

"Because of the difficulty in ensuring adequate comparability between study groups in an observational study -both when the project is being designed or upon review after the work has been published - we wonder whether non methodologically trained researchers can identify when potential selection bias or other biases more common with observational studies have occurred"

The complex nature of some health promotion interventions and the fact that they are sometimes evaluated using observational designs suggests that the assessment of quality presents particular challenges to reviewers that perhaps might not be as acute in other areas in health. Does this mean that systematic reviewers in health promotion require advanced training? Does it dissuade people from seeking training, and from doing systematic reviews in the first place? Are there any other specific challenges that face trainees in systematic reviewing health promotion? How are they overcome?

If sufficient capacity is to be available to meet the increasing demand for systematic reviews of the effectiveness of health promotion then it is important to investigate all the issues raised above in order to make recommendations for improvement, where appropriate. It is necessary to put these questions to systematic reviewers themselves, and also to providers of training and support who would be able to discuss their experiences and to offer views on the way forward. Therefore, objectives for this research are:

Research objective:

10. How do reviewers learn to do systematic reviews of health promotion?

- Which learning strategies are considered most successful?
- What are the barriers, to and facilitators of, learning?
- What are people's experiences of receiving training?

And:

Research objective:

11. What are reviewer's experiences of helping others to learn systematic reviewing?

- What forms of training and support are provided?
- What issues and topics are covered?
- What have been the challenges and successes in providing training and support?

1.9 Summary of research objectives

To re-iterate:

The overall aims of this study are:

1) To assess current approaches to, and general views on, systematic reviews of the effectiveness of health promotion; specifically the process of quality assessment of included evidence.
2) To identify areas of consensus and dissension around the choice of techniques, methods and criteria employed.

The research objectives and specific questions of the study are:

1. To assess current views on the strengths and weaknesses of systematic reviews of health promotion

2. To assess the challenges reviewers have faced when doing systematic reviews of health promotion:

- How have these challenges been dealt with?
- With what success?

3. To assess the extent to which systematic reviews of health promotion tend to assess the quality of included studies:

- What are the barriers to, and facilitators of, quality assessment?

4. To assess the extent to which quality assessment is conducted and reported in a 'systematic' manner:

- Do systematic reviews of health promotion tend to apply the same set of criteria to each study?
- Do systematic reviews of health promotion tend to single some studies out for criticism over others?
- Do systematic reviews tend to criticise studies for specific methodological flaws without having formally appraised them?

5. To assess how systematic reviews of health promotion use quality judgement:

- Do the findings and conclusions of systematic reviews reflect the strengths and weaknesses of the included studies?
- If so, by which methods? (e.g. quality thresholds; quality weighting, etc)
- Is there consensus on the most appropriate method?

6. To assess the criteria that systematic reviews of health promotion use to assess the quality of included evidence:

- Which criteria are used?
- Why have these criteria been chosen?
- Do these criteria address acknowledged threats to internal validity?

7. To assess whether there is consensus on the criteria by which health promotion evaluations should be assessed in systematic reviews.

8. To assess the extent to which systematic reviews of health promotion assess the external validity of included studies:

- For what purpose do systematic reviews of health promotion assess external validity?

9. To assess which types of people commonly participate in the production of systematic reviews of health promotion:

- Who does reviews (e.g. academics, health and other professionals, lay people), and what is their rationale for doing them?
- Who performs quality assessment in systematic reviews? (e.g. people who specialise in producing systematic reviews; people who specialise in the topic area being reviewed; combinations of these)
- To what extent are systematic reviews the product of collaborative teams? What are the advantages and disadvantages of collaborative team working?

10. How do reviewers learn to do systematic reviews of health promotion?

- Which learning strategies are considered most successful?

- What are the barriers, to and facilitators of, learning?
- What are people's experiences of receiving training?

11. What are reviewer's experiences of helping others to learn systematic reviewing?
- What forms of training and support are provided?
- What issues and topics are covered?
- What have been the challenges and successes in providing training and support?

The aims and objectives will be met by two sequential stages of research:

Stage 1) A systematic methodological mapping of a sample of health promotion systematic reviews to assess current approaches to, and consensus/dissension on, techniques, methods and criteria to assess quality.

Stage 2) A series of interviews with systematic reviewers in health promotion to assess and explore consensus/dissension over techniques, methods and criteria.

Table 3 shows at which stage each of the research objectives will be investigated. Some objectives are investigated by only one stage, whilst some are investigated in both stages.

Table 3 – Research objectives and corresponding stages of the research

Stages of the research	**Research objectives addressed**
Stage 1	3, 4, 5, 6, 7, 8, 9
Stage 2	1, 2, 3, 6, 7, 9, 10, 11

1.10 Chapter summary

This chapter has set the rationale for this study through a detailed review of the literature and proposal of a number of research objectives for this study. The next chapter provides an overview of the methodological framework used to meet the research objectives.

▪

Chapter 2 – Overall methodological framework

Chapter outline

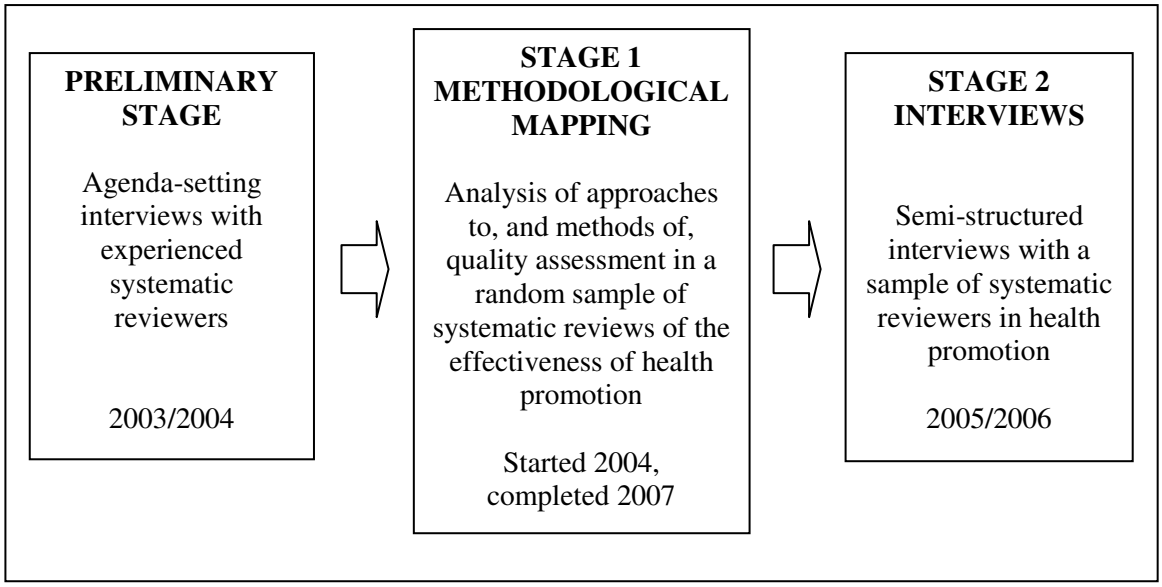
The aim of this chapter is to describe and justify the overall methodological framework used in this investigation. The first section provides a rationale for this framework. It then puts the methodology into context by discussing how it fits into the wider discipline of methodological research in health and health promotion.

2.1 Rationale for the methodological framework

2.1.1 Stages of the research

The research had three stages, intended to be sequential (Figure 4).

Figure 4 - Diagrammatic representation of methodological framework



The preliminary stage comprised a short agenda-setting exercise in which qualitative interviews were conducted with a small sample of systematic review experts (n=6). The purpose of the exercise was to seek the views of these experts on key issues for this research to investigate (in addition to those issues identified from the literature), and on the draft data extraction instrument to be used in Stage 1. The methods used in this exercise are described briefly in Appendix 1.

Stage 1 was a methodological mapping of systematic reviews of health promotion to identify and explain trends in methods and approaches to quality assessment (The methods used in this

stage are described in Chapter 3). Stage 2 comprised semi-structured interviews with systematic reviewers to investigate the trends and themes identified in Stage 1 (The methods used in this stage are described in Chapter 5).

The stages were originally designed to be sequential, with some of the (largely) quantitative findings from Stage 1 to be explored further in the more qualitative Stage 2. However, in reality there was some overlap between the two stages. Stage 1 began in early 2004 and was intended to be completed before Stage 2 started in late 2005. However, Stage 2 started slightly earlier than originally planned to take advantage of the 2005 Cochrane Colloquium which, because of its focus on health promotion and public health, was used to sample interviewees (as explained in more detail in Chapter 5, Section 5.4). Stage 1 therefore continued in the background during 2005/6 and was completed in early 2008.

2.1.2 Determining consensus

A central issue for this research to investigate is the notion of consensus in the methods to conduct systematic reviews in health promotion. There are a number of ways that consensus can be investigated. Commonly used methods include Delphi exercise (Dalkey and Helmer, 1963) the nominal group technique (Delbecq and Van de Ven, 1971), and the consensus development conference (Fink *et al*, 1984). These methods share the aim of reaching consensus on a given topic through seeking the opinion of experts in the field, using a variety of techniques (Jones and Hunter, 1995). In health care these methods have been used in the development of clinical guidelines (Black *et al*, 1999; Hutchings and Raine *et al*, 2006), and to set research priorities (Shepherd *et al* 2007a). Whilst there are published examples of Delphi exercises to develop criteria for the quality assessment of RCTs, these were not conducted within the context of health promotion (Sindhu *et al,* 1997; Verhagen *et al*, 1998).

A consensus-setting exercise would be more appropriate in this study had the aim been to *create* consensus. However, before this can be done the *extent* to which consensus already exists needs to be assessed. As the literature reviewed in Chapter 1 shows, few studies have attempted to assess this. The aim here is to assess the degree to which consensus exists, to understand causes of dissension, and identify ways to resolve it. This research can therefore be viewed as a precursor to a consensus-setting exercise. It is for this reason a framework was chosen that allows these issues to be explored using a multi method approach.

2.1.3 A multi method approach

The two contrasting main stages of research in this study were chosen to compliment each other. Some of the research objectives lent themselves to methodological mapping whilst others could only be fully investigated through interviewing systematic reviewers. At first glance this could be considered to be 'mixed methods' research, a popular and appealing approach used in a number of disciplines, including health (Harden and Thomas, 2005; Lingard *et al*, 2008).

Tashakkori and Teddlie (2003:711) define mixed methods as a:

"*Collection or analysis of both quantitative and qualitative data in a single study in which the data are collected concurrently or sequentially, are given a priority, and involve integration of the data at one or more stages in the process of research…when strategies derived from qualitative and quantitative methods are used within a single project"*

An example of mixed methods in evidence-based health would be qualitative process evaluation interviews with recipients of an intervention in an RCT (Protheroe *et al*, 2007). Morse (2003:190) is careful to distinguish between mixed method and multi method research. In mixed methods research one incorporates strategies that do not normally form a part of a particular research method in such a way that is congruent with the theory and principles of the core method. So, in the example above the qualitative process evaluation (not usually considered an integral feature of the quantitative RCT design) is incorporated into the RCT (the core method) to shed light on the outcomes (although, as mentioned in Chapter 1, RCTs are increasingly being designed with integral process evaluations). In multi method research the emphasis is upon one or more discrete types of research brought together on an equal footing, and is defined as:

"The conduct of two or more research methods, each conducted rigorously and complete in itself, in one project. The results are then triangulated to form a comprehensive whole"

In this study stages 1 (largely quantitative) and 2 (predominantly qualitative) are the discrete types of research that are used in complimentary fashion to meet the study's aims. Morse (2003) comments that multi method designs can be used inductively and deductively, though not both equally. The epistemology of this study is generally inductive, seeking to discover and explore, rather than deductive, aiming to test an *a priori* hypothesis.

The use of mixed / multi methods reflects an attempt to overcome the long-standing and much discussed 'paradigm wars' between quantitative and qualitative research across the sciences (Broom and Willis, 2007; Bryman, 2007; Guba and Lincoln, 2005; Hammersley, 1992; Oakley, 1999). It shares similarities with the technique of triangulation in that it aims to produce greater insight than would be gained by a single research method (Denzin, 1978). In this research four of the 11 research objectives were investigated by both research stages, thus using a triangulatory approach. The strength of this research, therefore, is that it is not reliant solely on one methodological paradigm (Cohen *et al*, 2007).

2.2 The context of this study: methodological research in evidence-based health

This study classifies itself as methodological research, that is, investigation into the methods used to conduct research which in this case is systematic reviewing. The last 10 to 15 years has seen an increase in publication of methodological research studies in health, designed to develop and improve the methodology of evidence synthesis and evaluation. Much of the impetus has been in recognition of the fact that methodological research has traditionally possessed few well-defined tools and processes analogous to those available for substantive research (Lilford *et al*, 2001). Studies have addressed methodological issues as diverse as literature searching, statistical synthesis of study results, and methods of assessing the degree of bias associated with different study designs / study methods. The latter is sometimes referred to as meta-epidemiology, and notable examples of these have been cited in Chapter 1 (e.g. Deeks *et al*, 2003; Schulz *et al*, 1995). I myself have a long-standing interest in methodology in health research and have contributed to a number of publications on methodological issues in evidence synthesis (Harden *et al*, 2004; Oliver *et al*, 2008; Shepherd and Harden, 2003; Shepherd *et al*, 2003; Shepherd *et al,* 2007a). As mentioned in the Introduction to this thesis, these projects have provided the impetus for the current research.

Empirical investigation into evaluation methodology and evidence synthesis is not new, however. The proliferation of methodological research in recent years rests on bedrock laid down over the decades by evaluators and social scientists (Campbell and Stanley, 1966; Cooper 1989; Cook and Campbell, 1979; Hedges and Cooper, 1994).

Much of the recent methodological research in health has taken place at the tertiary level, that is, analysis of secondary data (i.e. systematic reviews). Table 4 outlines a schema of different levels of evaluation (proposed by this study). This schema puts the mapping exercise in Stage 1 of this investigation into context (i.e. at the tertiary level).

Table 4 - Schema of different 'levels' of evaluation

Study Level	Study type
Primary	Primary evaluation of the effectiveness of an intervention (e.g. RCT)
Secondary	Systematic review of the effectiveness of interventions (e.g. of RCTs)
Tertiary	Systematic overview of systematic reviews (e.g. a mapping of *methods* used in, and *results* of, systematic reviews)

2.3 Chapter summary

This chapter has provided a brief overview of the methodological framework used in this study, one that combines both qualitative and quantitative data collection and analysis. The context within which this research is located has been discussed with reference to key texts on research methodology and methodological research in evidence based-health. As discussed, the research was conducted over two key stages, of which the first – the methodological mapping of systematic reviews of the effectiveness of health promotion – is described in the next chapter.

Chapter 3 - Methods for Stage 1: Methodological mapping of systematic reviews

Chapter outline

The aim of this chapter is to describe and justify the methods used in the first stage of the project, the methodological mapping of a sample of systematic reviews of the effectiveness of health promotion. Figure 5 (adapted from Figure 2 in Chapter 2) illustrates the methodological framework for the study as a whole, and highlights the sub-sections of this chapter and the methods discussed in each. First, the initial development of the draft data extraction instrument is described, and comments by systematic reviewers on the instrument as part of the agenda-setting exercise discussed in Chapter 3 are then summarised. The chapter then describes how the instrument was piloted and subsequently modified. A rationale for the sampling strategy is then presented and the chapter concludes by describing how the finalised instrument was applied to the reviews and how the results were analysed.

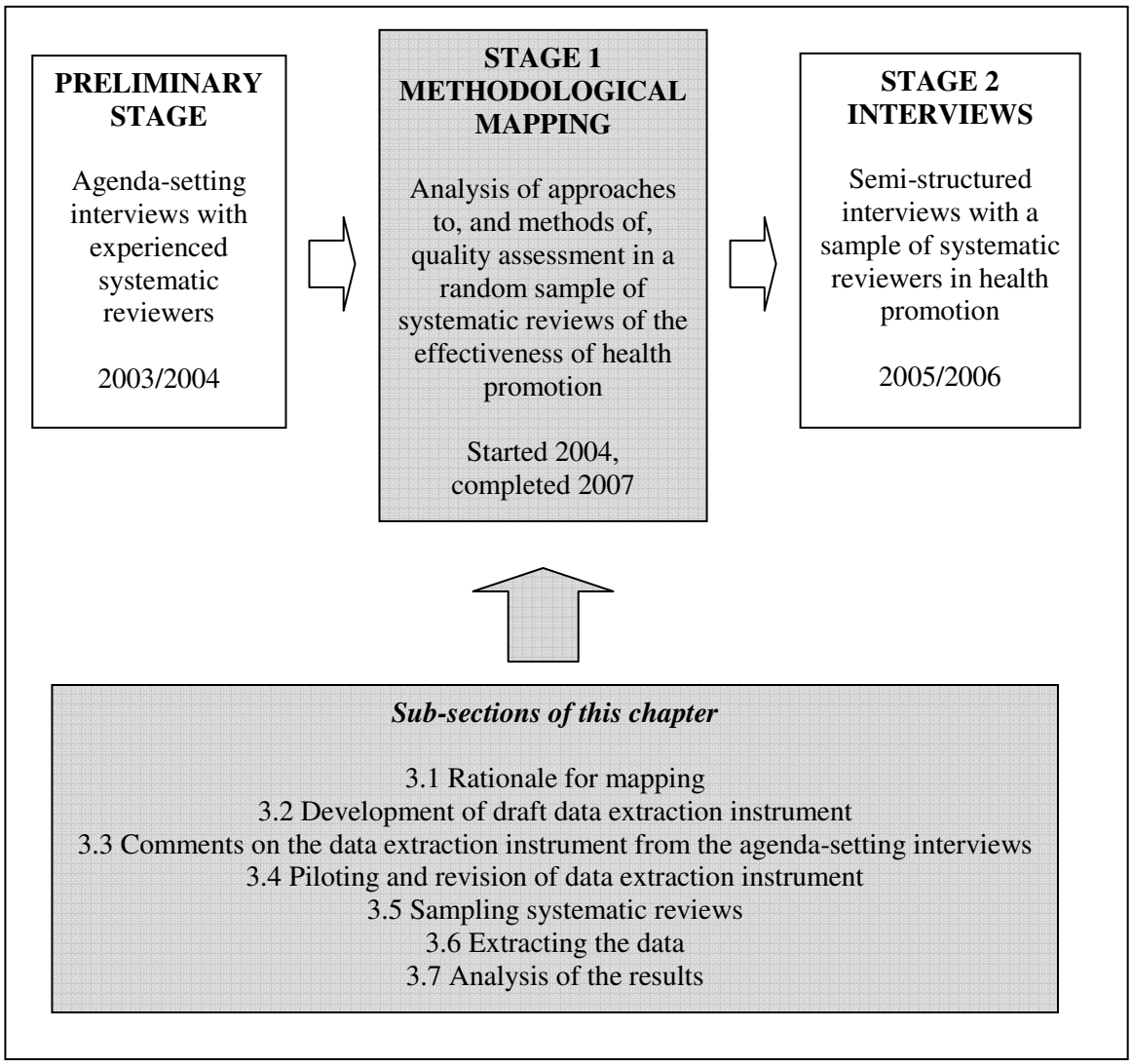
3.1 Rationale for mapping

In Chapter 1 a systematic mapping exercise was proposed to meet specific research objectives of this study. In this section a more detailed rationale for this exercise is given, with reference to published examples of what is becoming an increasingly used tool in evidence-based health.

In this study 'mapping' is defined as a systematic search for research studies (in this case systematic reviews) and the description of their key methodological characteristics using both qualitative and quantitative methods. Mapping can identify gaps where further research is needed, highlight particular areas where the volume of research is high and therefore where a systematic review would be useful, and bring to light potential flaws in the evidence where further, more rigorous, research is needed (for examples of a range of different mapping methodologies see the following: Coren and Bates, 2006; Doyle *et al*, 2005; Ellis *et al*, 2003; Ellis and Grey, 2004; Swann *et al*, 2003; Hawkins and Law, 2005; Katz *et al*, 2003; Shepherd *et al,* 2007a). Mapping can also be used to chart the methods used in research studies in order to identify trends and themes in methodology (Bryman, 2006; Paolucci El Dib, 2005).

In this study methodological mapping was used because its objective, systematic and transparent nature meant it was particularly appropriate to chart the methods used by systematic reviews of health promotion.

Figure 5 - Sub-sections of this chapter, and how they relate to Stage 1 and the study in general



3.2 Development of draft data extraction instrument

A draft data extraction instrument was devised during November 2003 (the finalised version is in Appendix 2). In developing the data extraction instrument the same principles employed in designing a questionnaire or interview schedule were adopted. That is, it was to be as reliable and valid as possible (Oppenheim, 1992). The instrument was designed to be semi-structured in nature, mostly comprising questions with pre-coded response categories plus some open-ended questions. The questions themselves were based on the research objectives identified in Chapter 1. The draft instrument comprised 41 questions (the finalised version comprised 50) and was structured around five themes each dealing with different aspects of a systematic review's methodology, as outlined in Table 5.

Table 5 - Data extraction instrument: key sections, themes and issues

Section	Theme	Issues covered	Relevant research objective(s)*
A	General details about the review	• scope, topic area, type of intervention	None – context setting
B	General details about the quality assessment process	• general approach to quality assessment used • how criteria were applied • how the results of the review reflect the strengths and weaknesses of included studies	3, 5, 9
C	Questions about the criteria to assess internal validity	• study design criteria • specific methodological attributes of study designs • justification for choice of criteria	7, 6, 4
D	Questions about the quality assessment instrument	• how it was devised, structured and validated	6
E	Questions about the criteria to assess external validity	• purpose for which external validity issues are addressed • specific aspects of the intervention examined • the study participants examined	8

* Please refer to Chapter 1 (Section 1.9) for the full list of research objectives

Two of the questions were adapted from an existing EPPI-Centre data extraction instrument routinely used in their systematic reviews (Peersman *et al*, 1997). These were general questions about the topic area and type of intervention (questions A4 and A5) and contain a large number of pre-coded response categories, which have been used in a number of EPPI-Centre systematic

reviews. It would have been inappropriate to duplicate available resources that could be used in this investigation.

The intention was to refine the draft instrument in two ways:

(i) To seek views on its content and structure through unstructured interviews with a small sample of experienced systematic reviewers in health promotion (i.e. agenda-setting interviews in the preliminary stage).

(ii) To conduct a pilot exercise by applying it to a sample of systematic reviews and making modifications as necessary.

3.3 Comments on the data extraction instrument from the agenda-setting interviews

As discussed in Chapter 2, one of the objectives of the agenda-setting interviews with the six systematic review experts was to elicit their comments on the draft data extraction instrument. Their key comments and the implications for the instrument are summarised briefly below.

Their comments were generally positive and criticisms were constructive. Many were of the opinion that the instrument required only minor changes. Some interviewees offered a number of highly specific suggestions for enhancing the instrument whilst others talked more generally about the study, making fewer comments directly about the instrument itself (as specified in Chapter 1). Their suggestions led me to make around 15 amendments to the instrument, mostly minor changes to existing questions.

3.4 Piloting and revision of the data extraction instrument

The second draft of the instrument was applied to a sample of five systematic reviews sampled randomly from the EPPI-Centre's online Database of Public Health Effectiveness Reviews (DoPHER).

The topics covered by the five reviews were:

- Pregnancy prevention amongst adolescents (Dicenso *et al*. 2002)
- Interventions to increase influenza immunization rates among high-risk populations (Sarnoff and Rundall, 1998)
- School based sex education (Silva *et al*, 2002)
- Prevention of sexually transmitted infections among young heterosexual men (Elwy *et al*, 2002)
- Wilderness challenge programs for delinquent youths (Wilson and Lipsey, 2000)

Each time the draft instrument was applied to a systematic review notes were made about additions or modifications necessary.

The diversity of topics, interventions and approaches to quality assessment as exemplified in this small sample of systematic reviews was an advantage for the purposes of piloting, as the aim of the data extraction instrument was to capture the multiplicity of approaches to systematic reviewing in health promotion. It should be noted, however, that these reviews were relatively explicit about the procedures for assessing quality. The standard of reporting was generally higher than in the reviews subsequently sampled in Stage 1 (see Chapter 4).

The instrument was finalised following comments made by the interviewees and from the piloting exercise. It was prepared for routine use by incorporating it into a specialist electronic database called EPPI-Reviewer (Thomas, 2002). EPPI-Reviewer is a specialist electronic database used by the EPPI-Centre to extract and store information from primary studies included in systematic reviews, including reviews that I had participated in (Oliver *et al*, 2008; Rees *et al*, 2006; Shepherd *et al,* 2006a). It can be accessed from any computer connected to the internet, via a web browser (see Figure 6), and is therefore easy to access, facilitating efficient fieldwork. I had used this database before to conduct systematic reviews whilst I worked at the EPPI-Centre, and thus I was already familiar with it. It was for these reasons I chose to use this database, as opposed to using a generic programme (such as Microsoft Access) which would have required time and expertise to design to the bespoke specifications of this project.

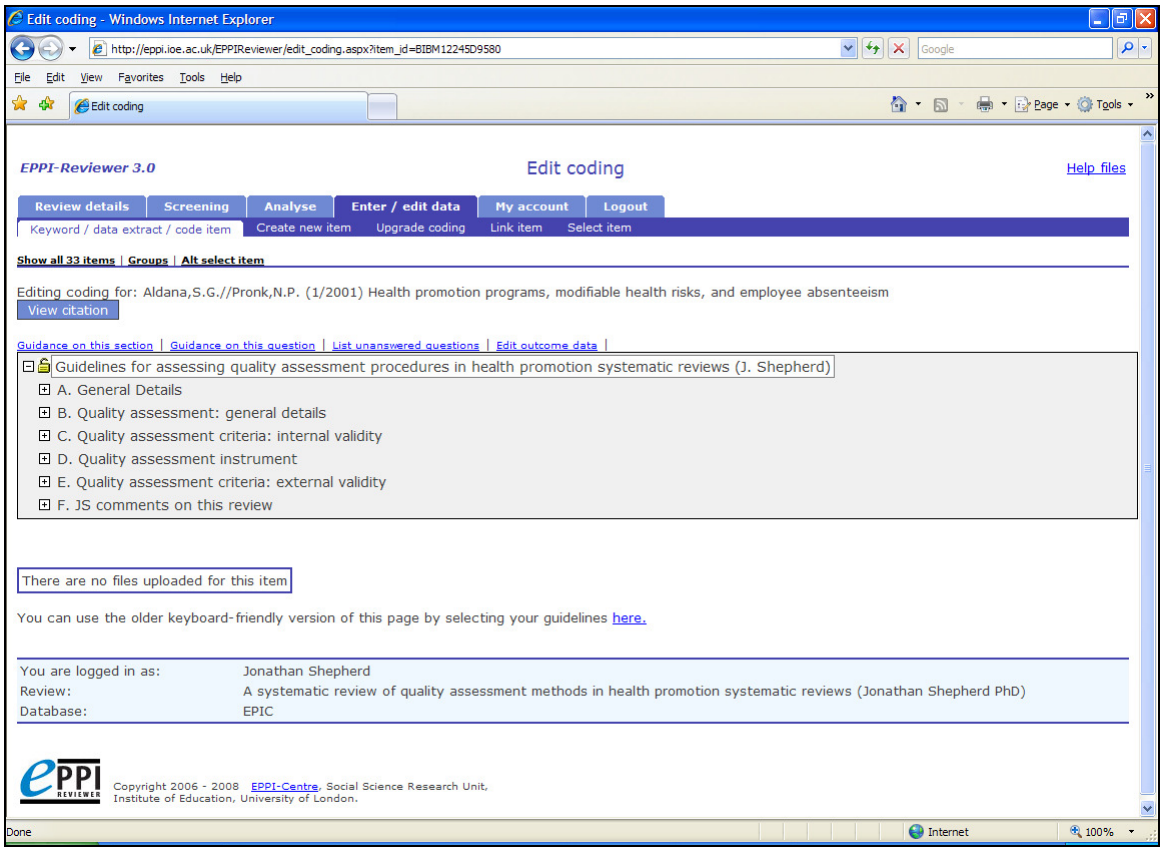
3.5 Sampling systematic reviews

In terms of sampling, the aim was to sample as many systematic reviews as was necessary until key themes and trends had been identified. In common with the pilot exercise, the DoPHER database was used as a source for sampling systematic reviews. The systematic methodology used to develop this database, and its comprehensive coverage, meant that it was a high quality source of evidence for this study to which no other credible alternatives were identified.

The Cochrane Database of Systematic Reviews (CDSR), one of the constituent databases of The Cochrane Library (Cochrane Collaboration, 2008) was considered as a potential alternative. However, this is restricted to reviews conducted by the Cochrane Collaboration, and would therefore produce a sample biased towards reviews conducted to the Collaboration's specification.

It would not be representative of the range of methodological approaches to systematic reviewing that this study aimed to assess. Also, identifying health promotion reviews on the CDSR has also been shown to be difficult due to unreliable indexing of keywords (Brunton *et al*, 2002).

Figure 6 - The finalised data extraction instrument in EPPI-Reviewer



Another alternative, the Database of Abstracts of Reviews of Effectiveness (DARE) produced by the CRD (Sowden and Glanville, 2006) was not used because it does not specialise in health promotion reviews

Methodological research has shown that indexing of health promotion studies in electronic bibliographic databases such as Medline is unreliable, making effective searching and screening time-consuming and laborious (Harden *et al*, 1999; Harden, 2001; Powell *et al*, 2005). Using DoPHER therefore saved me having to design, test, refine and run a search strategy to identify systematic reviews of health promotion, since this considerable amount of work had already been done during the creation of DoPHER. This represented efficient data collection and judicious use of my time and resources.

A random sample of 50 reviews was taken from DoPHER in November 2003, using the random numbers generator available in Microsoft Excel. A random sample was necessary to capture a sample of systematic reviews representative of different topic areas and methods. As Stage 1 took longer than anticipated (for reasons explained in Chapter 5, Section 5.4) a second, smaller, random sample (n=10) was performed in October 2007 to identify recently published reviews. This was to ensure that the overall sample reflected current as well as historical systematic review methods.

The inclusion criteria for this study were that the topic under review had to be within the realms of health promotion. The broad definition of health promotion used to determine inclusion of reviews in the DoPHER database itself accorded with the broad definition used in this study (Chapter 1, Section 1.3).

Preliminary analysis undertaken at around the 20th review indicated that, for many of the questions, similar trends and themes were emerging. It was therefore considered that saturation was approaching and continuing to the original target of 50 reviews would be unlikely to alter the results obtained at that point.

A total of 30 reviews were eventually included and data extracted (see Appendix 3 for a bibliography). These comprised 21 reviews from the original random sample of 50 in 2003, and nine from the random sample of ten taken in 2007. Note that the 21 reviews from the original sample of 50 were selected for data extraction randomly, rather than purposively (e.g. by date, or by topic area). This will have preserved the random nature of the sub-sample.

3.6 Extracting the data

In order to ensure consistency and objectivity this study followed the standard procedures for extracting data that would be employed in a systematic review itself (Higgins and Green, 2008; Petticrew and Roberts, 2006). Full reports of the systematic reviews were obtained from the EPPI-Centre, who have a hard copy of each of the reviews indexed in the DoPHER database. I read each review in turn and extracted data directly into EPPI-Reviewer.

Each of the questions with pre-coded categories were answered by placing ticks in the relevant categories, and qualitative detail was added in 'dialogue boxes' for each selected category to explain the rationale for choosing it. Dialogue boxes were also used for open-ended questions to capture qualitative data for which pre-coded categories were inappropriate (e.g. Question 'A.7 Authors' qualitative description of intervention'). As much detail was extracted as possible to

facilitate detailed analysis, to serve as a general aide-mémoire, and to prevent having to later refer back to the original publications (which can be time consuming and laborious). This was also necessary to reduce any bias associated with selective extraction of data (Petticrew and Roberts, 2006). Although the extracted data were not checked by a second reviewer for accuracy and inclusiveness (as would be standard practice in a systematic review), the transparent nature of the extraction means that this would at least be possible, if resources were available for such an exercise.

On average, each data extraction took between one to two days to complete, inclusive of reading the publication(s), extracting the data, and checking the finished data extraction for accuracy and reflecting on the fairness of interpretation.

3.7 Analysis of results

The principles underpinning the analysis were that it should be systematic and transparent to enable it to potentially be reproduced by others, a basic pre-requisite for all academic research (McBurney, 2001; Cohen *et al*, 2007; Robson, 2002). Although mapping exercises such as this have become more common, little methodological guidance has been published on how they should be conducted and analysed. Therefore, this study has used similar analytical techniques used by other peer-reviewed published mapping studies, such as Moja *et al* (2005) who, as mentioned in Chapter 1, assessed the comprehensives of quality assessment procedures in systematic reviews of health care.

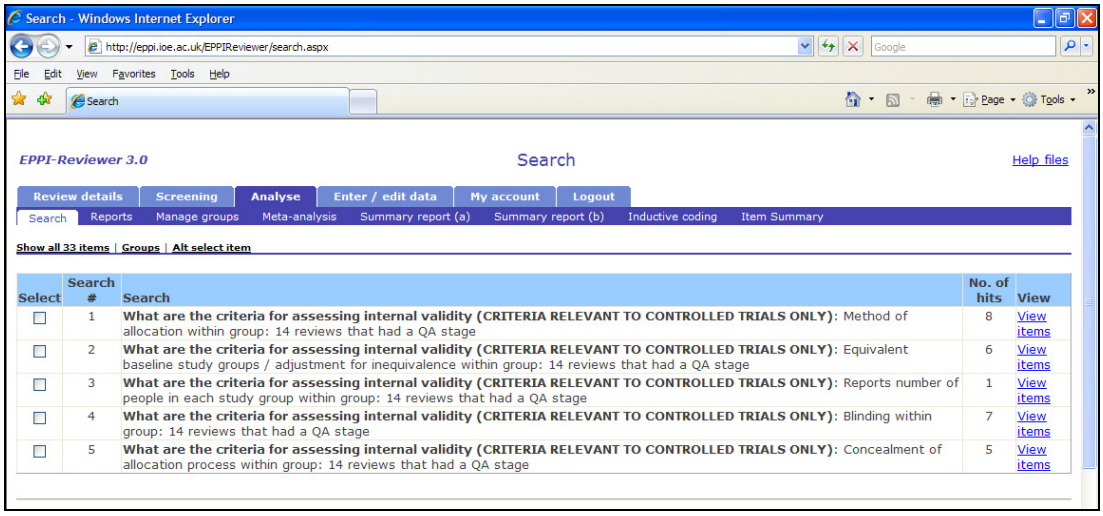
To some extent the analysis of a mapping exercise is a straightforward process and does not necessarily require sophisticated analytical methods. Rather, established analytical techniques used in other forms of investigation including survey research, interviewing, document analysis, and systematic reviewing itself, were applied. For example, Bowling (2002: 378) discusses the analysis of documentary evidence such as policy documents and official statistics. A form of content analysis, similar to that employed in the analysis of qualitative interview data, can be used to analyse documents:

"The systematic and objective identification, linking and counting of specified characteristics [that] can be carried out in order to compare categories and to make inferences from the data"

A similar content analysis style approach was used in this study to categorise the qualitative data extracted from the systematic reviews (see below).

In this study a combination of quantitative and qualitative analysis was undertaken, in accordance with the nature of the data extracted. Each of the six sub-sections of the data extraction instrument, and each question within the sub-sections, was analysed in turn. Frequencies for responses to the questions with pre-coded categorical responses were generated. This was done in EPPI-Reviewer using the 'search' function. Figure 7 is an illustration of a search for the frequency with which different quality criteria were employed in the systematic reviews. The number of reviews featuring each criterion is listed on the right hand side in the column 'No. of hits'. Clicking on the 'View items' hyperlink generates the bibliographic details of the reviews featuring each criterion. All of the searches were saved in EPPI-Reviewer to permit re-analysis as necessary. The frequencies generated were then presented in the results chapter (e.g. for the frequencies presented in Figure 7 see Table 11 in Chapter 4).

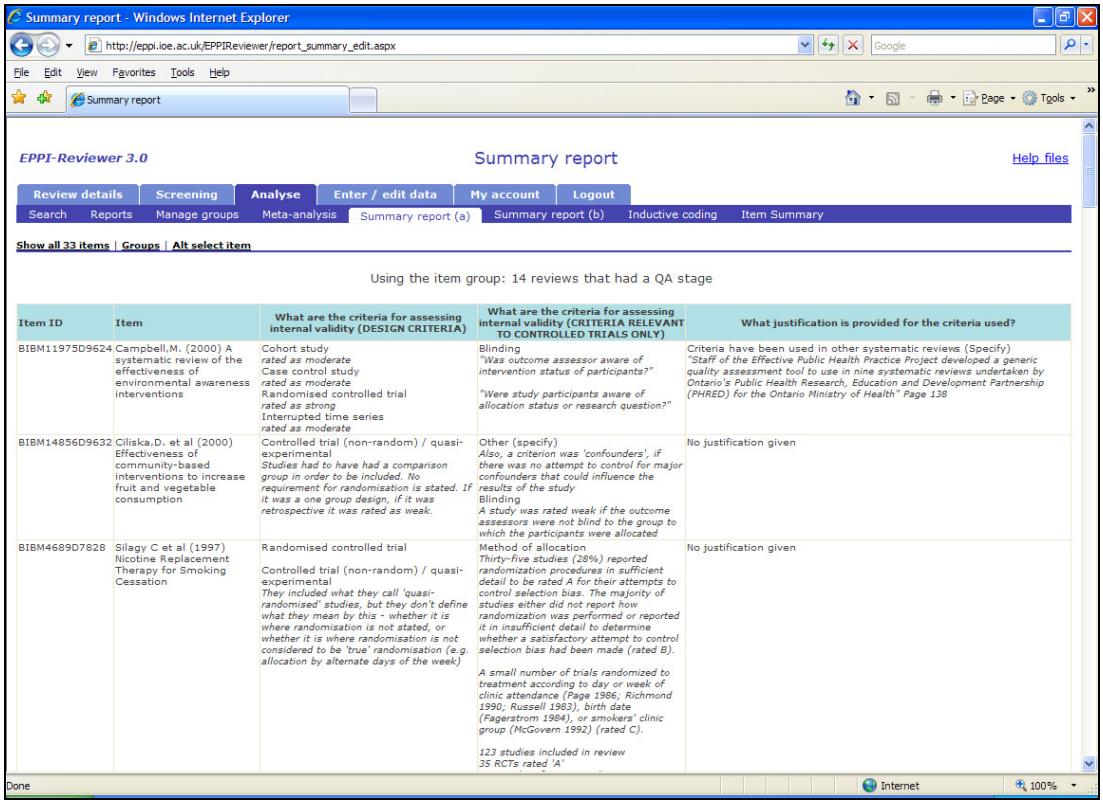
Figure 7 - Example of frequency tabulation in EPPI-Reviewer



The answers to the open-ended questions were a mixture of my observations on the reviews, and quotes from the review publications copied verbatim for reference. This qualitative detail was retrieved for analysis in EPPI-Reviewer using the 'Reports' function which automatically cross-tabulates the responses to a given number of questions. This saved the time usually needed to copy and paste qualitative text into tables by hand.

Figure 8 provides an illustration of one particular cross-tabulation. In this example three questions relating to internal validity were cross-tabulated, and the responses given for first three systematic reviews (under the column heading 'Item') are shown (the remainder of the reviews would be seen by scrolling down the screen). The text in italics is the qualitative detail entered during data extraction. This text was used to select illustrative examples in the analysis and presentation of the findings (Chapter 4).

Figure 8– Cross-tabulated data analysis in EPPI-Reviewer



3.8 Chapter summary

This chapter has described and justified the methods used in the first stage of data collection, the methodological mapping of systematic reviews of the effectiveness of health promotion. It has reported how a data extraction instrument was piloted and applied to a random sample of systematic reviews, and described how the data were analysed. The next chapter presents the results of this exercise.

Chapter 4 - Results of Stage 1: Methodological mapping of systematic reviews

Chapter outline

The chapter presents the results of the first stage of the research, the mapping of a random sample of 30 systematic reviews. It starts by presenting the key characteristics of the reviews to set the context. The extent to which the reviews assess methodological quality is described, in terms of which stages of the review process this is undertaken, and methods used to ensure the results of the review reflect the strengths and weaknesses of the evidence. The criteria used to assess quality in the reviews are presented, followed by an analysis of the extent to which there is a consensus over criteria. An analysis of the extent to which quality is assessed in a systematic manner is described, followed by an examination of the characteristics of the reviewers in terms of their academic / professional status. The chapter concludes with an investigation into the extent to which the reviews consider external validity, and for what purpose.

Recap: research objectives relevant to Stage 1

To re-iterate, seven of the 11 research objectives were relevant to this stage of the research:

3. To assess the extent to which systematic reviews of health promotion assess the quality of included studies:

- What are the barriers to, and facilitators of, quality assessment?

4. To assess the extent to which quality assessment is conducted and reported in a 'systematic' manner.

- Do systematic reviews of health promotion apply the same set of criteria to each study?
- Do systematic reviews of health promotion single some studies out for criticism over others?
- Do systematic reviews criticise studies for specific methodological flaws without having formally appraised them?

5. To assess how systematic reviews of health promotion make use of quality judgement:

- Do the findings and conclusions of systematic reviews reflect the strengths and weaknesses of the included studies?
- If so, by which methods? (e.g. quality thresholds; quality weighting, etc)
- Is there consensus on the most appropriate method?

6. To assess the criteria that systematic reviews of health promotion use to assess the quality of included evidence:

- Which criteria are used?
- Why have these criteria been chosen?
- Do these criteria address acknowledged threats to internal validity?

7. To assess whether there is consensus on the criteria by which health promotion evaluations should be assessed in systematic reviews.

8. To assess the extent to which systematic reviews of health promotion assess the external validity of included studies:

- For what purpose do systematic reviews of health promotion assess external validity?

9. To assess which types of people commonly participate in the production of systematic reviews of health promotion:

- Who does reviews (e.g. academics, health and other professionals, lay people), and what is their rationale for doing them?
- Who performs quality assessment in systematic reviews? (e.g. people who specialise in producing systematic reviews; people who specialise in the topic area being reviewed; combinations of these)
- To what extent are systematic reviews the product of collaborative teams? What are the advantages and disadvantages of collaborative team working?

4.1 Key characteristics of the systematic reviews sampled

This section sets the context for the chapter by briefly describing the key characteristics of the systematic reviews included in this investigation.

A total of 30 reviews were analysed (see Appendix 3 for a bibliography). Dates of publication ranged from pre-1980 (the earliest published in 1978) to 2006 (Table 4.1, Appendix 4). Half were published between 1995 and 2000 (n=15/30; 50%). The majority (n=23/30; 77%) were published in peer-reviewed journals (Table 4.2, Appendix 4). Five (17%) of the reviews were conducted for the Cochrane Collaboration.

The length of the reviews, and consequently the level of detail reported, varied. Due to the finite word limits imposed by academic journals, the majority of reviews provided limited details of

their methodology. This lack of detail restricted what could be extracted for this investigation, except where ancillary publications were available.

The most commonly reviewed topics included the prevention of sexually transmitted infections, pregnancy prevention, prevention or cessation of tobacco use, and the promotion of healthy eating and physical activity (Table 4.3, Appendix 4). The majority of reviews focused on one particular topic, or group of related topics.

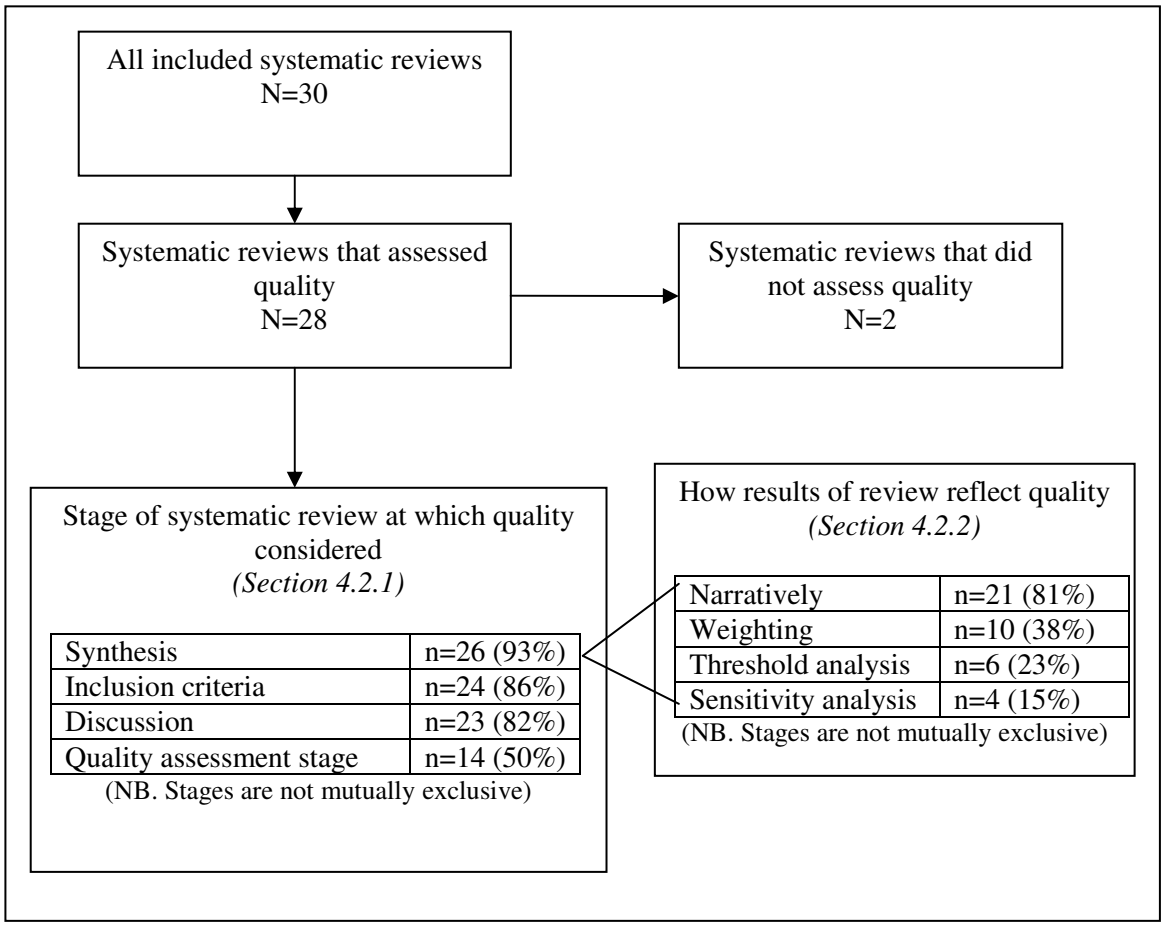
The most common type of intervention involved the provision of information and education, a feature of 21 (70%) of the reviews (Table 4.4, Appendix 4). Other common types of intervention included the provision of advice and counselling (n=14/30; 47%), and practical skill development (n=12/30; 40%). Only three (10%) of the reviews were classified as including studies that evaluated only one type of intervention. The majority of reviews, therefore, included studies in which the intervention comprised more than one type of activity (n=27/30; 90%).

4.2 To what extent do systematic reviews of health promotion assess quality?

Figure 9 illustrates the number of reviews classified as assessing quality. The vast majority of reviews assessed the quality of the included studies (n=28/30; 93%), with only two reviews (7%) not reporting any form of quality assessment. Figure 9 also shows how many reviews were classified as assessing quality at different stages of the review process, and the frequency with which different methods were used to consider quality within one of these stages, the 'synthesis' stage. [The purpose of the figure is to set the scene for the following sub-sections which describe, in greater detail, the different stages, and the different methods used. The tables presented in the remainder of this chapter are based on various sub-sets of systematic reviews, therefore Figure 9 can be seen as a 'reference point' for the chapter].

Of the 28 reviews that considered quality the extent to which they did so varied considerably. Some reviews paid particular attention to study quality. For example, two of the 28 reviews described themselves as methodological reviews, focusing more on the methodological quality of the included studies than their outcomes (Booth and Watters, 1994; Tingle *et al*, 2003). In another review assessment of methodological quality was stated as one of the objectives of the review (Huibers *et al*, 2003).

Figure 9 – Flowchart illustrating the proportion of reviews that assessed quality, and at what stage quality was considered



4.2.1 At which stage of the review was quality assessed?

Table 6 shows the stages of the systematic reviews at which study quality was assessed amongst the 28 reviews that were classified as having considered quality.

Quality was most often considered during the synthesis stage of a review (n=26/28; 93%). During this stage reviews often commented on the strengths and weaknesses of studies in relation to presenting their results (see Section 4.2.2 for details of the different methods used in the synthesis).

The second most common stage where quality was considered was during the process of assessing studies for inclusion / exclusion (n= 24/28; 86%). Typically, the reviews specified that, to be eligible for inclusion, a study must use particular evaluation design and/or certain methodological attributes must be present.

Table 6 - Stage of systematic review when quality is considered (sub-set of 28 reviews that assessed study quality)

Stage of review	Proportion of reviews N (%)
Synthesis	26 (93)
Inclusion / exclusion criteria	24 (86)
Discussion	23 (82)
Quality assessment stage	14 (50)

NB. Reviews could assess quality at more than one stage, hence total numbers exceed 28

During this process studies were not necessarily formally appraised in detail, but screened for the presence of certain methodological characteristics (discussed in greater detail later in this chapter, see Section 4.3.1.1).

The quality of included studies was also commonly considered by reviews when discussing their overall findings (n=23/28; 82%). In common with the standardised format of research reports, the latter sections of many of the systematic review reports discussed the implications of the overall findings, and reflected on the quality of the evidence base. The extent to which issues of quality were discussed varied from a couple of cursory sentences, to lengthy discussion spanning several pages. For example, the discussion section of the Cochrane review of psychosocial interventions delivered by general practitioners (Huibers *et al,* 2003: 8), contained a sub-section entitled 'methodological quality' in which the authors discussed in detail the strengths and weaknesses of the included studies, as summarised by this quote:

"In summary, despite fairly good methodological quality and positive findings of some studies, evidence for the effectiveness of psychosocial interventions by general practitioners does not exceed level 3 (limited or conflicting evidence), except for good evidence that problem-solving treatment is no less effective than antidepressant treatment for depression"

Comparatively fewer reviews reported conducting a quality assessment exercise (n=14/28; 50% reviews). In this investigation a review was classified as doing this if they mentioned applying quality assessment criteria to each included study. Those reviews that did report such a stage tended to do so as part of a data extraction exercise following the completion of literature searching and screening studies for inclusion.

4.2.1.1 Assessment of quality at multiple stages of the review

Of the 28 reviews that considered quality, 26 (93%) of them did so at more than one stage of the review. Table 7 outlines 15 different scenarios based on all of the different combinations of stages at which it is possible to assess quality (based on the four stages presented earlier in Table 6). Scenario D was the most common (n=11/28; 39%), whereby quality is assessed at study inclusion, as a quality assessment stage, and is integrated into the synthesis and discussion of study results.

Table 7 - Scenarios for stage (s) at which quality is assessed in the reviews (sub-set of 28 reviews that assessed study quality)

Scenario A (n=2)	**Scenario B (n=0)**	**Scenario C (n=1)**	**Scenario D (n=11)**
A) Inclusion criteria	A) Inclusion criteria	A) Inclusion criteria	A) Inclusion criteria
	B) QA stage	B) QA stage	B) QA stage
		C) Synthesis	C) Synthesis
			D) Discussion
Scenario E (n=1)	**Scenario F (n=2)**	**Scenario G (n=7)**	**Scenario H (n=0)**
A) Inclusion criteria	A) Inclusion criteria	A) Inclusion criteria	A) Inclusion
C) Synthesis	D) Discussion	C) Synthesis	B) QA stage
		D) Discussion	D) Discussion
Scenario I (n=0)	**Scenario J (n=0)**	**Scenario K (n=0)**	**Scenario L (n=2)**
B) QA stage	B) QA stage	B) QA stage	B) QA stage
	C) Synthesis	D) Discussion	C) Synthesis
			D) Discussion
Scenario M (n=1)	**Scenario (n=1)**	**Scenario (n=0)**	
C) Synthesis	C) Synthesis	D) Discussion	
	D) Discussion		

QA = quality assessment

Scenario G was the second most common, with seven reviews (25%) classified as considering quality when assessing studies for inclusion (A), when analysing results (C), when discussing the findings (D), but without reporting any kind of formal assessment of quality (B). Only one review considered quality at just one of the four stages (Cross *et al*, 1998).

4.2.2 How do the results of systematic reviews reflect the quality of the studies?

As reported earlier, 26 of the 28 reviews (93%) that considered quality did so at the 'synthesis' stage (Table 6). In these reviews the analysis of the results of the included studies took into account their strengths and weaknesses, so that potential biases were made explicit in the presentation of the findings. The reviews were classified as doing this in four different ways, as outlined in Table 8.

Table 8 – Methods used by reviews to consider quality during the synthesis of results (sub-set of 26 reviews that considered study quality at the 'synthesis' stage)

Method of considering quality within the synthesis	Proportion of reviews N (%)
Narratively	21 (81)
Weighting	10 (38)
Threshold analysis	6 (23)
Sensitivity analysis	4 (15)

NB. Reviews could use more than one method, hence numbers total numbers exceed 26

The most common method was a narrative description of the quality of the included studies in relation to their results (n=21/26; 81% reviews). These reviews commonly described the methodological quality of the evidence base in general. For example, Shults *et al* (2001: 75) in their review of the effectiveness of interventions to reduce alcohol impaired driving, commented that minimum legal drinking age (MLDA) laws were effective at preventing accidents, and were based on what they considered to be strong evidence:

"According to the Community Guide's rules of evidence, there is strong evidence that MLDA laws, particularly those that set the MLDA at age 21, are effective in preventing alcohol-related crashes and associated injuries"

Ten of the 26 reviews (36%) gave more weight to studies judged to be of better quality. For example, in the review of workplace nutrition and cholesterol interventions by Glanz *et al* (1996), all studies were tabulated according to a star rating, with the best quality studies receiving five stars.

Six of the reviews (23%) employed a threshold analysis, whereby only studies meeting a particular threshold of quality were included in the synthesis of results. For example, in a review

of the effectiveness of environmental awareness interventions by Campbell (2000), 65 studies met the inclusion criteria. Of these, only 14 were judged to be either 'strong' or 'moderate' in methodological quality, and were therefore eligible to be analysed in detail.

Four reviews (15%) conducted sensitivity analysis to explore how the removal or addition of studies judged to be of poorer quality influences the overall findings of the review. Three of these were Cochrane reviews and planned such analysis *a priori* in the protocol that all Cochrane reviews publish in advance. The fourth review, by Ogilvie *et al* (2004; 2005), decided to conduct sensitivity analysis upon completion of their review:

"When our review was complete, we also conducted a sensitivity analysis to examine what the content and findings of the review would have been if we had taken one of two extreme approaches to inclusion—either (a) by restricting the review to randomised controlled trials, or (b) by including all relevant studies. This sensitivity analysis was intended to answer two questions: were the conclusions of our review sensitive to the inclusion criteria, and could we have reached our conclusions more efficiently?"
(Ogilvie *et al*, 2005: 887)

4.2.2.1 Multiple methods of assessing quality in the synthesis

Table 9 outlines 15 different scenarios based on all of the different combinations of methods that can be used to consider quality within the synthesis stage of a review (based on the four methods presented earlier in Table 8).

Just under two-thirds of the reviews (n=15/26; 58%) employed only one of the four methods for addressing quality. Of these the most common was the narrative method (Scenario A), which was the sole approach used by 10 (38%) of the reviews, followed by threshold analysis (Scenario M; n=2/26; 8%), and by sensitivity analysis (Scenario O; n=1/26; 4%) and weighting (Scenario I; n=1/26; 4 %). The remaining 10 reviews (40%) employed multiple methods to address quality in the synthesis. Of these the most common was a combination of a narrative approach and giving the results of the better quality studies more weight (Scenario B, n=6/26; 23%). Another combination was a narrative approach in addition to threshold analysis (Scenario E, n=3/26; 12%). Only one review used all four methods to address quality within the synthesis (Scenario D, n=1/26; 4%). This was the aforementioned review of the effectiveness of promoting walking and cycling by Ogilvie *et al* (2004; 2005).

Table 9 - Scenarios for different methods for considering quality within the synthesis stage of a review (sub-set of 26 reviews that considered study quality at the 'synthesis' stage)

Scenario A (n=10)	**Scenario B (n=6)**	**Scenario C (n=0)**	**Scenario D (n=1)**
A) Narratively	A) Narratively	A) Narratively	A) Narratively
	B) Weighting	B) Weighting	B) Weighting
		C) Threshold analysis	C) Threshold analysis
			D) Sensitivity analysis
Scenario E (n=3)	**Scenario F (n=0)**	**Scenario G (n=0)**	**Scenario H (n=1)**
A) Narratively	A) Narratively	A) Narratively	A) Narratively
C) Threshold analysis	D) Sensitivity analysis	C) Threshold analysis	B) Weighting
		D) Sensitivity analysis	D) Sensitivity analysis
Scenario I (n=1)	**Scenario J (n=0)**	**Scenario K (n=1)**	**Scenario L (n=0)**
B) Weighting	B) Weighting	B) Weighting	B) Weighting
	C) Threshold analysis	D) Sensitivity analysis	C) Threshold analysis
			D) Sensitivity analysis
Scenario M (n=2)	**Scenario N (n=0)**	**Scenario O (n=1)**	
C) Threshold analysis	C) Threshold analysis	D) Sensitivity analysis	
	D) Sensitivity analysis		

The reviews generally did not provide a rationale for the approach or combination of approaches used to consider quality within the synthesis. One exception was the Cochrane review of psychosocial interventions delivered by general practitioners by Huibers *et al* (2003), described earlier. From the primary publication of this review it was considered that a narrative approach had been followed, and that studies had also been weighted according to a rating of quality. An ancillary publication (van Tulder *et al*, 1997) described the range of different methods that could be used to consider quality within the synthesis, and the strengths and weaknesses of each, with reference to empirical literature. This demonstrated that the authors had carefully considered the merits of the various approaches in their choice of method.

4.2.3 What are the barriers and facilitators to the process of quality assessment?

Only three reviews (10%) reported barriers to the process of quality assessment (Dyson *et al,* 2005: Hillsdon *et al,* 2005; Huibers *et al,* 2003). All mentioned that poor reporting of their included studies hampered attempts to make a thorough judgement, and all endeavoured to overcome this limitation by contacting authors of studies where necessary to elicit missing information. In general this was reported to be a successful process. Huibers *et al* (2003)

reported that seven out of eight authors contacted replied with the information necessary for them to decide whether or not they were eligible for inclusion. Hillsdon *et al* (2005) wrote to, and received information from, the authors of five of the 17 studies included in the review. Data from a sixth study were reported to be unavailable and consequently that it was excluded from the synthesis of results.

Dyson *et al* (2005: 4) were less successful in their attempts to elicit information from study authors. Their inclusion criteria specified that studies must have used a RCT design, yet this information was not clear from all the potentially relevant studies:

"We contacted authors to clarify or obtain relevant details of individual studies, particularly to request details of their randomisation processes"

Fifty studies were excluded from their review for various reasons, leaving seven to be included. At least five studies were not able to be included because of missing information. In three of these cases it was not clear whether studies had been randomised adequately, despite efforts to contact the author.

The review by Powell *et al* (2004) also reported the intention to contact authors in cases of missing information, but did not need to as no studies were eligible for inclusion in the review.

Kirby (2006: 11), in contrast to the other reviews, routinely contacted authors of all included studies. The purpose of this exercise appeared to be to enable those authors to comment on the completeness and accuracy of the data that had been extracted by the systematic reviewers, and the fairness of the quality judgement made.

"The revised templates were then sent to the original study authors for verification. The authors of 73 out of 83 studies reviewed the summaries, and suggested minor changes that were incorporated. The remaining authors did not respond despite subsequent requests. (All 83 of these one-page summaries are available from the authors upon request)"

The remaining reviews generally did not report the existence of any barriers or facilitators.

4.3 Which criteria do systematic reviews of health promotion use to assess quality?

4.3.1 Criteria used

4.3.1.1 Criteria for evaluation designs

Table 10 shows, in rank order, which designs were permitted by the 24 reviews that specified evaluation design as an inclusion criterion. Randomised and non-randomised / quasi experimental designs were permitted in the vast majority of reviews (n= 23/24; 96% and n=19/24; 79%, respectively). Four (17%) reviews permitted inclusion only of RCTs, three of these being Cochrane reviews. Eight (33%) reviews permitted RCTs in addition to non-randomised controlled / quasi-experimental designs. The remainder of the reviews included a range of evaluation designs, including cohort studies and case control studies. Six (25%) of the reviews had what this study classes as an 'open-door' policy, and did not restrict inclusion criteria to any particular evaluation design.

Table 10 – Evaluation designs permitted for inclusion in reviews (sub-set of 24 reviews that specified evaluation design as an inclusion criterion)

Evaluation design	Proportion of reviews permitting evaluation design N (%)
Randomised controlled trial	23 (96)
Controlled trial (non-random) / quasi-experimental	19 (79)
One group pre and post	10 (42)
Post test only, 1 group	9 (38)
Post test only, >1 group	9 (38)
Cohort study	9 (38)
Case control study	8 (33)
Interrupted time series	7 (29)
Case series	7 (29)
Case study	7 (29)
Other (e.g. "Independent cross-sectional design")	1 (4)

NB. Reviews could permit more than one evaluation design, hence total numbers exceed 24

Similar results were observed when the analysis of evaluation designs was restricted to the 12 reviews which considered study quality as an inclusion criterion and also reported a quality assessment stage (data not shown). That is, reviews which reported a quality assessment stage

did not differ from those which did not, in terms of the proportions permitting each evaluation design. The following sub-sections present the criteria employed by the sub-set of 14 reviews that reported a quality assessment exercise.

4.3.1.2 Criteria specific to controlled trials

Table 11 presents a list of criteria that apply only to controlled trials (including randomised controlled trials) and the proportion of systematic reviews that featured each criterion in their assessment of quality (n=14 reviews reporting a quality assessment exercise).

Table 11 - Proportion of systematic reviews that featured criteria specific to controlled trials (n= sub-set of 14 reviews that reported a quality assessment exercise)

Criterion	**Number of reviews featuring criterion N (%)**
Method of allocation to study groups	8 (57)
Blinding	7 (50)
Baseline study groups / adjustment for in-equivalence within group	6 (43)
Concealment of allocation to study groups	5 (36)
Number of people in each study group reported	1 (7)

NB. Reviews could permit more than one criterion hence total numbers exceed 14

Blinding

Reviews which assessed 'blinding' (n=7/14; 50%) examined whether the different people involved in an evaluation were aware of the intervention recipients' group assignment: those measuring study outcomes (n=7/14; 50%, 'detection bias'), those receiving the intervention themselves ('performance bias', n= 2/14; 14%), and those providing the intervention (n=1/14; 7%). A common theme was the belief that concealing which intervention a participant had been assigned to from them was not as feasible for health promotion as other kinds of intervention. For example, Hillsdon *et al* (2005: 5) commented:

"We did not rate studies on whether participants were blind to their allocation to intervention or control groups. This would not be appropriate for studies of this type, as it would be impossible to blind participants to a physical activity intervention"

Dyson *et al* (2005: 5), in their review of interventions to promote the uptake of breastfeeding, were mindful not to unfairly criticise individual studies for failing to mask the identity of the intervention. Rather, they acknowledged that the problem was inherent in the evidence base for this kind of intervention:

"Given that there are genuine pragmatic considerations when delivering and evaluating breastfeeding promotion interventions, the ability to reduce performance bias is limited and this should be recognised as an inherent weakness of this particular type of evidence base rather than of the particular studies included in this review"

Despite the difficulties in masking the identity of health promotion interventions from their recipients, Dyson *et al* (2005: 6) considered that this was achievable for some health promotion interventions. They gave an example of the intervention employed by one of the studies in their review, analogous to a drug intervention:

"The only study which was considered to have adequately addressed potential sources of performance bias was the evaluation of a breastfeeding promotion pack compared to a commercial formula pack (Howard, 2000), a study which was able to maintain blinding of both participants and providers through the use of sealed, similarly designed, packs more comparable with the use of a placebo and treatment in a therapeutic trial"

Dyson *et al* (2005) also suggested that certain outcomes are less amenable to bias, and therefore, whether or not the assessor was blind to participant assignment is less important.

Other criteria

Reviews which assessed 'method of allocation to study groups' (n=8/14; 57%) commonly examined whether or not study participants had been assigned to groups in a random, quasi-random or non-random fashion. Furthermore, where randomisation had been reported it was assessed whether the method used could truly be considered random.

Reviews which assessed 'baseline study groups / adjustment for in-equivalence within group' (n=6/14; 43%) examined how similar study groups were at the start of the study (i.e. at 'baseline') in terms of social and demographic factors, and factors associated with the intended outcomes of the intervention. The purpose was to assess whether or not selection bias had occurred, and hence whether the results may be confounded due to differences in groups. If

baseline differences were reported by the studies some of the reviews then appraised whether or not any statistical adjustments had been made to compensate for bias.

The five reviews that assessed 'concealment of allocation to study groups' were all Cochrane systematic reviews. All Cochrane reviews use software (Review Manager) in which the criterion of allocation concealment is a standardised feature. By default these reviews therefore appraised studies using the rating system of A, B, C or D (A -adequate; B - unclear; C - inadequate or D – not used). Only one of the reviews discussed the importance of this criterion (Huibers *et al*, 2003).

4.3.1.3 Criteria applicable to all evaluation designs

Table 12 presents a list of criteria that apply to all evaluation designs and the proportion of systematic reviews that featured each criterion in their assessment of quality (n=14 reviews reporting a quality assessment exercise).

Table 12 – Proportion of systematic reviews that featured criteria applicable to all evaluation designs (n=sub-set of 14 reviews that reported a quality assessment exercise)

Criterion	**Number of reviews featuring criterion N (%)**
Attrition / Loss to follow-up discussed	12 (86)
Validity and reliability of data collection instruments/methods	9 (64)
Validity and reliability of data analysis methods	8 (57)
Sample size	7 (50)
Length of follow-up	5 (36)
Outcome measures / All outcomes reported on	5 (36)
Contamination / co-intervention	4 (26)
Pre- and post-intervention data provided	3 (21)
Hawthorne effect / testing effect	1 (7)
Clearly defined aims	1 (7)
Temporal trends	1 (7)
Publication status	1 (7)

NB. Reviews could permit more than one criterion hence total numbers exceed 14

The most commonly cited criterion was whether studies took into account the effects of attrition and loss to follow-up of participants (n=12/14; 86% reviews). The reviews assessed a number of different (non-mutually exclusive) aspects of attrition, with the most common being whether attrition rates had been reported and whether the volume of drop-out could be considered a significant threat of bias (n=11/12; 92%). The level of attrition considered acceptable by the reviews varied, but in general did not exceed 40% of the total study population. Other aspects of attrition assessed included whether or not the data analysis accounted for the effects of participant drop-out (n=4/12; 33%); whether reasons were given for why participants dropped-out (n=1/12; 8%); and whether the rate of attrition was similar between study groups (n=1/12; 8%).

The second most common criterion was the validity and reliability of the instruments and methods used to measure the effectiveness of the interventions (n=9/14; 64%). Within this criterion there were five mutually exclusive sub-categories that the reviews fell into. Five of the reviews assessed whether the instruments (e.g. questionnaires; interviews; biochemical tests, etc) had been validated and were known to be reliable. Two reviews considered the limitations of certain kinds of data collection instruments. The remaining two reviews assessed the validity of wider aspects of data collection, including whether confidentially had been assured to participants when self-reporting their alcohol consumption, and whether pre and post intervention data in an uncontrolled evaluation were collected at comparable times of the year to reduce seasonal confounders.

The validity and reliability of data analysis methods was also a common criterion, as featured in eight reviews (57%). Within this criterion there were six non-mutually exclusive categories the reviews were classified by. Four of the reviews assessed whether studies had employed an 'intention to treat / intervene' analysis to compensate for study attrition, cross-over between study groups, and/or missing data. Two reviews assessed whether the unit of analysis in an evaluation was the same as the unit of allocation to study groups, a criterion only applicable to controlled trials. This was particularly important where clusters of people were allocated to study groups whereby analysing outcomes at the unit of the individual could confound results, as commented by one of the reviews (Fletcher and Rake, 1998). Other aspects of data analysis assessed included: 'proper use of statistical methods', defined as the inclusion of some level of practical significance in addition to statistical significance (n=1/8; 12.5%), whether the point estimate and measure of variability (e.g. standard deviation / standard error) were presented for the primary outcome measure (n=1/8; 12.5%), and use and reports of statistical methods controlling for design effects (n=1/8; 12.5%).

Seven (50%) reviews assessed sample size as a criterion. There were three mutually exclusive aspects of this criterion. Four reviews criticised studies for having 'small' sample sizes, but only Ogilvie *et al* (2004; 2005) qualified this with a threshold, specifying that there should be a minimum of 100 people in each study group. Two reviews required studies to have reported a sample size calculation sufficient to be able to detect a statistically significant difference between study groups. The remaining review only required that the sample size was described for each study group.

The length of follow-up of outcome measurement was included as a criterion in five (36%) reviews. The reviews appraised whether the duration was sufficient to capture all of the intended outcomes of the intervention evaluated. Two of the five reviews (40%) specified a minimum duration that they considered acceptable, varying from three to six months. Three of the five reviews (60%) commented on whether the duration had been short or long, but without qualifying what they meant by this. One of these also assessed whether timing of the outcome assessment was the same in all of the study groups.

The phenomenon of contamination from co-interventions was a criterion assessed in four reviews. For example, Booth and Watters (1994) reviewed the effectiveness of office-based risk reduction strategies for injecting drug users, and suggested that in at least three of the included studies contamination occurred from nearby outreach activities running at the same time.

4.3.1.4 Rationale for the criteria used

Table 13 presents a classification of the justifications given for the choice of quality assessment criteria and the proportion of reviews that cited each of them.

Ten of the reviews reported at least one justification, and four provided no rationale for their choice of quality assessment criteria. The most common justification, as reported by seven reviews, was that the criteria they used had been featured in a previous published systematic review. Slightly fewer reviews mentioned that their choice of criteria was based upon recommendations from guidelines on systematic reviewing. The least common justification was that the criteria had been developed with reference to empirical texts on evaluation methodology.

As Table 13 shows, the reviews could report more than one justification. Four reviews reported only one of the justifications, three reviews reported two of the justifications, and the other three reported all three justifications (data not shown).

Table 13 - Proportion of reviews citing different types of justification for the use of quality criteria (sub-set of 14 reviews that reported a quality assessment exercise)

Justification given	Proportion of reviews N (%)
Justification given	10 (71)
Criteria are recommended by systematic reviews	7 (70)
Criteria are recommended by systematic review methodology guidelines	6 (60)
Criteria supported by empirical evidence on protection against bias	4 (40)
No justification given	4 (29)

NB. Reviews could cite more than one justification, hence why numbers exceed 14

An example of the latter was the review by Tingle *et al* (2003), which cited what appear to be two empirical texts as underpinning their choice of quality assessment criteria. One was by Cohen (1988) on the topic of statistical power analysis in the behavioural sciences, whilst the other was a text on the impact of attrition on the internal validity of smoking prevention interventions by Biglan *et al* (1987). They also mention that their criteria have been used in other published systematic reviews.

Criteria are recommended by systematic reviews

As mentioned in Table 13, seven reviews reported that the quality assessment criteria they employed had been used in previous published systematic reviews, including the authors' own systematic reviews in some cases (e.g. Ogilvie *et al*, 2004; 2005). Five of the seven reviews (71%) also mentioned that the criteria they had employed had been adapted, rather than just reproduced, from criteria used in previous reviews. Citations to the criteria that they adapted are reported in Appendix 4 (Table 4.5). The citations were diverse, and none were used in more than one review.

Some of the systematic reviews suggested that the criteria they had adapted were high profile. For example, Huibers *et al* (2003: 4) who used the 'Maastricht-Amsterdam Criteria List (MACL)' criteria noted:

"The Maastricht-Amsterdam Criteria List (MACL) includes all criteria of other prominent quality scales like the Jadad List (Jadad 1996) and the Delphi List (Verhagen 1998)"

Criteria are recommended by systematic review methodology guidelines

The Cochrane Handbook was the most commonly cited set of guidelines, mentioned by four reviews (Table 4.6, Appendix 4). It is worth noting that the recommendations in the Handbook are based upon empirical investigations into a number of dimensions of methodological quality, and that the Handbook itself cites a number of empirical texts on bias and evaluation methodology. The other five guidelines were cited by one review each. They included guidelines issued by Cochrane review groups for use in systematic reviews of a particular topic, and guidelines from government funded health research organisations such as the US Agency for Health Care Policy and Research, and the UK CRD.

Three of the reviews cited more than one set of guidelines. For example Powell *et al* (2004), in their Cochrane review of visual acuity screening, cited section 6 of the Cochrane Handbook (the section that discusses methods of quality assessment) in addition to guidelines issued by the Cochrane Eyes and Vision Group. The former source provided recommendations for methods and criteria for use in systematic reviews in general, whilst the latter contained recommendations for criteria specific to that topic area.

Criteria supported by empirical evidence on protection against bias

A range of citations were by the four reviews classified as justifying their use of criteria with reference to empirical research (Table, 4.7, Appendix 4). With the exception of Shults *et al* (2001), the reviews cited more than one text in their discussion of study quality, but note that none of the texts were cited by more than one review. The texts varied in chronology, from the works of Campbell and Stanley in the 1960s, to relatively recent studies such as the investigation into the influence of various dimensions of methodological quality on the effects of controlled clinical trials, such as Schulz *et al* (1995). In terms of an example, Booth and Watters (1994: 1516) summarised the strengths and weaknesses of each of the different evaluation designs included in their review with reference to two texts. Appraising the 'one-group pre-test- post-test' design they commented:

"This paradigm cannot control for most threats to internal validity, including history, maturation, testing, instrumentation, interaction effects, and regression toward the mean (Cook and Campbell (1979)"

Discussing the 'pre-test post-test control group' design they remarked:

"In this design, participants are randomly assigned to experimental or control conditions. Randomisation can help assure initial group comparability…this design provides the best single means for increasing confidence in causal inference (Cook and Campbell, 1979)" (Booth and Watters, 1994: 1521)

The review by Huibers *et al* (2003) did not elaborate in detail on the justification for their choice of criteria, but referenced the source of their quality assessment instrument (van Tulder *et al* (1997)). The van Tulder paper itself provided an overview of the empirical evidence for dimensions of methodological quality known to influence study effects. They remarked that (at the time) there was paucity of empirical evidence on bias, and that the evidence that is available yields inconsistent findings:

"Currently, there is still limited empirical evidence of a relation between specific methodologic criteria and bias. Some authors have reported that inadequate concealment of treatment allocation is associated with larger effect sizes (Chalmers et al, 1983; Schulz et al, 1995) whereas others reported a bias in the opposite direction (Colditz et al, 1989)"
(van Tulder *et al*, 1997: 2323)

4.4 To what extent is there consensus over quality assessment criteria?

The results presented in the previous sections show that the majority of systematic reviews in this investigation included experimental evaluation designs, primarily RCTs. A relatively smaller proportion of reviews (less than half) have, in addition, permitted inclusion of observational studies, such as uncontrolled evaluations (refer back to Table 10). Experimental designs are therefore favoured, in terms of the types of study that are included in systematic reviews. In order to further assess whether this constitutes a consensus, comments on the strengths and weaknesses of different evaluation designs made by the authors of the systematic reviews were systematically extracted and classified. These comments are indicative of the authors' views on evaluation design (at the time of publication). The following sub-sections present examples of comments on the strengths and weaknesses of RCTs and observational studies.

4.4.1 Comments on the strengths of RCTs

Of the 28 (93%) reviews that considered the quality of included studies (refer back to Figure 9), 16 (57%) commented on the strengths of RCTs. These reviews varied in the degree of praise given. In some cases support for RCTs was implicit, such as the review by Hurtsi and Sjoden

(1997: 109) which implied that experimental studies would provide them with more confidence in the findings:

"Only studies with experimental/quasi-experimental design were included in order to facilitate conclusions about the effectiveness of the different programmes. All but two studies used random assignment of the participants to the study groups"

In other reviews support for RCTs was more explicit. Stout and Rivara (1989: 377) commented that the studies included in their systematic review of the effectiveness of sex education did not measure up to the high standards associated with RCTs:

"The 'gold' standard for any intervention, whether it be a drug trial or a community health program, is a randomised clinical trial. The studies reviewed here are far from this standard"

4.4.1.1 Recommendations for future evaluation designs

The vast majority of the reviews made recommendations for future research based on their findings (n=26/28; 93%). Of these 26 reviews, 14 (54%) advocated the use of RCTs and experimental designs.

Kirby (2006: 8) in his review of the impact of sex and HIV education interventions for young people remarked:

"Evaluations can and should use randomised experimental designs"

Such recommendations tended to reflect the general tone of the articles with respect to the merits of different evaluation designs and hierarchies of evidence. That is, they were consistent with the quality assessment criteria used in those reviews, and the justifications given for adopting those criteria (e.g. empirical texts, guidelines of conducting systematic reviews, and references to other systematic reviews - as discussed earlier in Section 4.3.1.4)

Less commonly the recommendations appeared to be in response to the poor quality of the evidence base reviewed. For example, in a review of the effectiveness of interventions to improve awareness of environmental hazards the authors remarked:

"This systematic review cannot comment on the effectiveness of mass distribution of printed materials as nearly all evaluation studies (11/13) of this intervention type were of weak

quality… high quality evaluation research is required to determine if common strategies such as tax bill inserts or mass distribution of pamphlets, posters or factsheets to community locations or targeted mailing lists are effective"
(Campbell, 2000: 142)

Some reviews went beyond merely advocating particular types of evaluation design, and made practical suggestions about how recommendations could be implemented. For example, Stout and Rivara (1989: 378) in their review of sex education in schools suggested potential locations that would be appropriate for an RCT:

"An appropriate evaluation of sex education programs should not be impossible to accomplish. Large cities with rates of teenage sexual activity such as Baltimore, Philadelphia, or Washington DC would serve as ideal sites for such a study and would provide a large enough sample to avoid a significant type II error. The intervention could be randomly assigned and evaluated in a prospective fashion"

Kirby (2006: 48) acknowledged that there may be practical challenges to mounting experimental evaluations, particularly in developing countries, but urged perseverance:

"One of the largest and most rigorous studies in the entire world was conducted in Mwanza, Tanzania. Other studies have implemented rigorous evaluation designs in developing countries. It is not always easy, but it can be done"

Foxcroft (1997: 536) proposed that rigorous evaluation of health promotion should be part of the infrastructure of a project:

"More funds should be targeted towards well designed evaluation studies...we suggest that rigorous evaluations should be defined and built into projects as a condition of funding, so that a cycle of high quality evidence-based practice is developed and maintained"

4.4.2 Comments on the weaknesses of RCTs

Eight of the 28 reviews (26%) commented on the weaknesses of RCTs. Despite the overall support for RCTs in the reviews included in this investigation, some of them conceded that, if not conducted properly, they could be flawed. Criticisms of RCTs fell into two categories (i) that RCTs can be poorly conducted, negating the merits inherent in the design (n=5/8; 63%);

and (ii) that RCTs may not be feasible in certain circumstances (n=4/8; 50%) (NB. One review fell into both categories, hence why total numbers exceed eight).

4.4.2.1 RCTs can be poorly conducted

Booth and Watters (1994: 1520) suggested that one of the greatest merits of this design is its ability to create study groups that are comparable at the initiation of the study, but acknowledged that this can be compromised by an uneven distribution of participant drop-out:

"Randomisation can help assure initial group comparability, post-test comparability, however, may be affected by differential attrition between groups, particularly when participation in the experimental condition is more demanding than in the control condition, as is often the case"

Booth and Watters (1994: 1520) then specify what they consider to be the pre-requisites for a sound RCT:

"This design provides the best single means for increasing confidence in causal inference as long as the following four conditions are met: 1) groups are equivalent at the pre-test; 2) experimental and control sessions, along with pre-tests and subsequent post-tests, are run simultaneously; 3) post-test data are analysed for all subjects, not just those who participated in the interventions and 4) differential attrition between conditions is not a factor"

Similarly, Glanz *et al* (1996) noted that attrition compromised the results of RCTs included in their review.

Other criticisms of the conduct and reporting of RCTs mentioned by the reviews included poor description of the method of random allocation to study groups (Ciliska *et al,* 2000); inability of cluster RCTs to recruit and randomise sufficient numbers of clusters to ensure an even distribution between study groups (Shults *et al,* 2001); and, also in relation to cluster RCTs, erroneously analysing the impact of the intervention on individual participants instead of the clusters (Fletcher and Rake, 1998).

4.4.2.2 RCTs may not always be feasible

One review suggested that RCTs are not applicable to interventions that promote health by changes in policy, or via the mass media:

"There is also a need to expand evaluation techniques beyond randomised controlled trials since not all public awareness interventions are conducive to this evaluation type, particularly those that deal with media awareness, advocacy or policy based activities"
(Campbell, 2000: 143)

Similarly, the methodological guide used to underpin the review by Shults *et al* (2001) and others in the series of reviews by The US Task Force on Community Preventive Services, commented:

"However, randomization is sometimes not feasible or ethical in population based research…"
(Briss *et al*, 2000: 41)

This comment was echoed by another review which added that RCTs of area-wide and community interventions are expensive:

"However, a randomised research design is extremely difficult and costly to implement in community or population based studies"
(Dishman and Buckman, 1996: 714)

Hurtsi and Sjoden (1997: 109) discussed one of the studies included in their review, in which randomisation to study groups appeared to contradict the aims of the intervention, which was to encourage families and their friends to exercise together in a community leisure facility:

"Although necessary to secure the internal validity, random assignment may also create problems. This was noted by Baranowski et al who reported a decreasing interest by the participants after the seventh week in the centre-based family intervention study. The families had been randomly assigned to the intervention condition and some of their close friends or relatives were assigned to the control group. Thus, a natural, already existing support mechanism was interfered with"

4.4.3 Comments on observational studies

The preceding sections have illustrated that, in general, randomised and experimental evaluation designs are considered to be the most rigorous form of evaluation by systematic reviewers. The corollary to this is that observational evaluation designs are, by default, considered inferior. The comments made by the reviewers generally endorse this.

However, in a few instances observational studies were mentioned in a positive context. For example, Booth and Watters (1994: 1522) put things into perspective by commenting on some of the first evaluations of HIV prevention, rapidly conducted during the formative stages of the AIDS epidemic. They remarked that the need for timely evidence for the effectiveness of interventions was hampered by a lack of resources:

"It is understandable why the first responses to the AIDS epidemic did not emphasize research and evaluation. In most communities, the threat of disease transmission was such that immediate action was required. The costs associated with conducting rigorous evaluation studies far exceeded both the time and resources available, given the potential benefits of specific interventions. Consequently, many of these research efforts represented the best effort possible under the circumstances"

Booth and Watters (1994: 1522) exercised a more lenient view of these studies, making allowances for the circumstances in which they were conducted:

"The fact that these early studies were imperfect should, we think, in no way detract from their utility"

A similar view was taken by Aldana and Pronk (2001: 44) who identified only two experimental evaluations amongst the 43 studies included in their review of health promotion interventions in the workplace. The majority of studies used quasi-experimental or correlational designs. They commented that some evaluators had to contend with challenging circumstances:

"Most of these studies were conducted in the real world setting, meaning that the researchers had to make the best of difficult research conditions. Most studies used non-random control groups or they statistically controlled for confounding variables and used large sample sizes"

Despite their reservations about the ability of such designs to adequately demonstrate causality in effect, Aldana and Pronk (2001: 44) were still able to draw some conclusions:

"Nevertheless, researchers have demonstrated some clear associations, especially in the areas of stress, obesity, fitness and health promotion program participation, and multiple risk factors"

4.5 To what extent is quality assessment conducted and reported in a 'systematic' manner?

Instances where the reviews appeared to be unclear, or inconsistent in their approach to quality assessment were recorded. Fourteen (50%) of the 28 reviews that assessed quality were classified as being ambiguous or inconsistent in their approach. This section presents some key examples from these reviews.

The review of interventions to facilitate employment in disabled people by Bambra *et al* (2005) was an example of a review which reported applying quality assessment criteria, but did not report what the criteria were. It was mentioned that their criteria were adapted from the literature (citations provided), however, it could not be discerned which study attributes were appraised. Studies that did not meet the criteria were not permitted for inclusion in the final review. All that is reported is that:

"Studies of any type with substantive flaws were excluded from the final review of the evidence" (Bambra *et al*, 2005: 1908)

It is therefore unclear on what grounds studies were excluded.

The review of risk-reduction interventions targeting injecting drug users by Booth and Watters (1994) appeared to have undertaken some degree of quality assessment, but it was not explicit how it had been conducted. The evaluation design of each of the 27 included studies appeared to have been classified, with a discussion given of the strengths and weaknesses of each design in turn. However, it is not clear whether each of the 27 studies were appraised in a systematic manner, and whether any particular criteria were applied, although the authors cite the seminal work by Campbell and Stanley (1966) and Cook and Campbell (1979) as underpinning their critique. The authors did not necessarily consider their review to be a detailed critical appraisal of the evidence, but do mention that their critique was systematic:

"This is not intended to be a state of the art literature review, but a critical comparison of published reports, selected from generally accepted and easily accessed sources. Our purpose is not to point to flaws, but to assess the degree to which these studies permit causal inference, and to provide a systematic approach to assessing this research portfolio"
(Booth and Watters, 1994: 1516)

An example of a review which provided a fairly lengthy methodological critique of the evidence, but reported vague details about procedures for assessing quality was that by Glanz *et*

al (1996). A rating system was used to appraise the quality of the study design used, with randomised studies receiving more stars (i.e. a higher rating) than others. The rating scale did not report specific methodological attributes to be appraised (as acknowledged by the authors), yet shortcomings of the studies, such as attrition and biases associated with self-reported outcomes are mentioned in their discussion. It is not clear whether the authors systematically assessed each study in terms of these attributes, or whether it is a general observation on the quality of the evidence.

Likewise, Kirby (2006) provided an extensive critique of the evidence for the effectiveness of sex and HIV education interventions, despite not reporting any details of how quality was assessed. A particular issue singled out for criticism was whether or not sample sizes were adequate to permit outcomes to be statistically significant. He commented that sample sizes need to be large enough to enable significant differences in rare outcomes to be identified:

"Given that only five of the 13 studies measuring impact on pregnancy had sample sizes greater than 2,000 and given that only two of the 10 studies measuring impact on STD rates had sample sizes greater than 2,000, the failure of these results to provide many statistically significant results does not necessarily mean that the programs did not have a programmatically meaningful impact on pregnancy or STD rates" (Kirby, 2006: 18)

However, no justification, statistical or otherwise, was provided for the sample sizes suggested. Likewise, the Cochrane review by Dyson *et al* (2005) also commented on whether or not studies had sufficient sample sizes, yet this was not reported to be one of their quality assessment criteria.

Four reviews, including Kirby (2006), criticised studies for employing inappropriate outcome measures despite this not being reported as a quality criterion.

4.6 Who participates in the production of systematic reviews of health promotion?

4.6.1 Types of people who conduct systematic reviews of health promotion

Table 14 reports a classification of the type of people who participated in the production of the systematic reviews, in terms of their academic, professional or public status. This information was only explicitly reported in a minority of reviews. Classification was therefore performed by examining the authors' academic and / or professional affiliations reported in the review publication, where supplied.

Table 14 - Classification of the types of people involved in conducting the systematic reviews (all 30 included reviews)

Type of person involved in systematic reviewing	Proportion of reviews N (%)
Researcher	20 (66)
Practitioner	6 (20)
Lay person / consumer	1 (3)
Student	1 (3)
Policy specialist	1 (3)
Not stated / unclear	9 (30)

NB. More than one type of person could participate in each review, hence why total numbers exceed 30. The unit of classification is the review

Researchers were the most common type of person involved in systematic reviewing, as classified in two-thirds of the reviews. They were commonly affiliated with universities, other academic institutions, or research organisations and were therefore classified as having a research role, amongst their other duties.

Less commonly involved in reviewing were practitioners, as classified in only six reviews. In four of these six reviews the practitioners were clinicians who conducted research whilst also practising medicine. Involvement of lay people, students and policy specialists was minimal. It was not possible to classify which types of people were involved in nine of the reviews as no author affiliations were reported.

Six (20%) of the reviews were classified as being conducted by more than one type of person. Typically researchers and practitioners collaborated to conduct these reviews. One notable example of multi-disciplinary collaboration was the review of interventions to reduce alcohol-impaired driving by Shults *et al* (2001). Three different teams were convened to produce the review. The 'co-ordination team' drafted the conceptual framework for the reviews, managed the data collection and review process, and drafted evidence tables, summaries of the evidence, and the reports. The 'consultation team' reviewed and commented on materials developed by the coordination team and set priorities for the reviews. The 'abstraction team' collected and recorded data from studies for possible inclusion in the systematic reviews. Collectively the teams included experts in the field of crash and injury prevention with backgrounds in medicine, public health, economics, health promotion intervention design and implementation, health education, health policy, and epidemiology.

In eight reviews it was possible to ascertain who participated in specific tasks in the production of the review. These reviews provided a breakdown of which authors contributed to which task, usually in a sub-section towards the end of the publication entitled 'Contribution of the authors'. All five Cochrane systematic reviews reported a breakdown, in accordance with the standard format used by the Cochrane Collaboration for reporting reviews. To illustrate, below is a typical example of the information reported in one such Cochrane review:

'Marcus Huibers (MH) and Anna Beurskens (AB) identified and selected all studies. In case of doubt, they consulted Gijs Bleijenberg (GB) for advice on the selection of studies. AB and GB assessed the methodological quality of selected studies and performed the data extraction'
(Huibers *et al*, 2003: 21)

The eight reviews were analysed to identify which types of people conducted quality assessment. Researchers were classified as having participated in all eight reviews, practitioners were classified in three reviews, whilst policy specialists were classified in just one review.

4.6.2 Specialist backgrounds of systematic reviewers

Table 15 classifies the reviews in terms of whether the people conducting them specialised in the topic area under review, or whether they specialised in conducting systematic reviews (not necessarily in one specialist topic area).

Table 15 - Specialist background of systematic reviewer (all 30 included reviews)

Background	**Proportion of reviews N (%)**
Topic specialist	18 (60%)
Systematic review specialist	7 (23%)
Can't tell/ not stated	8 (27%)

NB. People from more than one specialist background could be classified as participating in each review, hence total numbers exceed 30

The largest proportion classified was topic specialists, defined in this investigation as people (primarily the reviews' authors) who appeared to have expertise in the particular topic under review, as discerned from examination of their professional and/ or academic affiliation, their publications (as listed in the review's bibliography), and any other information provided in the publication. Internet searches were also conducted to identify and examine their biographies

(e.g. via search engines such as www.Google.com) in order to obtain further detail on their background.

An example of a review conducted by topic specialists was the review of interventions to promote the initiation of breastfeeding by Dyson *et al* (2005). The three authors of this review were classified based on biographies identified through internet searching. All three were affiliated with the Mother and Infant Research Unit at the University of York, and the unit's 2004 annual report gave detailed information on their professional and academic interests, publications and projects. The report confirmed that collectively they had a strong academic and / or professional background in child and maternal health, and nutrition.

As reported in Table 15, seven reviews were classified as being conducted by systematic review specialists. This classification, like that of topic specialist, was based on information provided in the publications, and from internet searches I conducted specifically to ascertain the background of the authors. An example was the review of the effectiveness of community-based interventions to increase fruit and vegetable consumption by Ciliska *et al* (2000). This review was conducted as part of a series of public health reviews by the Effective Public Health Practice Project in Ontario, Canada. The reviews conducted in the series cover a diverse range of topics, and although the teams convened to conduct the reviews included specialists in the topic area, they also included project staff whose primary role is to routinely conduct systematic reviews.

Six (20%) of the reviews were classified as having been conducted by both 'systematic review specialists' and 'topic specialists' (not shown in the table). For example, the aforementioned review by Shults *et al* (2001), comprised a multi-disciplinary team of people including 'methodologic experts in systematic reviews' (Zaza *et al* 2001: 24) as well as experts in the prevention of road traffic injuries.

Note that the classifications systematic review specialists and topic specialists in this study are mutually exclusive, such that someone skilled in systematic reviewing would commonly review a variety of topics, as opposed to reviewing one particular area. Similarly, someone with expert knowledge of a particular topic area would not necessarily be expected to be an accomplished systematic reviewer. However, there was some evidence to suggest a 'cross-over' in specialist expertise. For example, the biography of the systematic reviewer Lisa Dyson, mentioned above (Mother and Infant Research Unit, 2004), refers only to her expertise of nutrition in the context of child and maternal health, but also that she had become increasingly active in systematically reviewing this topic area (Mother and Infant Research Unit, 2004).

4.6.3 Training for systematic reviewers

None of the systematic reviews reported whether or not the people conducting them had received any training. A couple of authors did, however, comment briefly about training and support. Tingle *et al* (2003) reported that students who participated in the production of their review were taking part in a University 'Educational Research Methodology Program'. It is presumed that the program covered the theory and practice of systematic reviewing although few details of the curriculum are reported. Foxcroft *et al* (1997) mentioned that their systematic review was monitored closely by staff from the Centre for Reviews and Dissemination, University of York. No further details are given.

4.7 To what extent do systematic reviews of health promotion consider external validity?

Table 16 reports a classification of the purpose for which they systematic reviews assessed issues of external validity.

Table 16 – For what purpose does the review address external validity? (all 30 included reviews)

Purpose	Proportion of reviews N (%)
To facilitate generalisability / replicability	24 (80)
To explain results	17 (56)
To assess the quality of the intervention	11 (37)

NB. Reviews could be classified as assessing external validity for more than one purpose, hence total numbers exceed 30

The majority of the reviews (n=18/30; 60%) were classified as assessing external validity for more than one reason. In some reviews it was explicit why external validity issues were assessed, as evident from the review's objectives, from the data extraction and quality assessment criteria employed, and from descriptive data on the characteristics of interventions and study populations presented (e.g. Shults *et al*, 2001; Hillsdon *et al*, 2005). However, in other reviews the purpose was not always explicitly stated. Reviews were therefore classified in this investigation by a systematic assessment of their objectives (where given), as well as from a systematic assessment of comments made in the review publications.

The following sub-sections describe in greater detail each of the categories in Table 16.

4.7.1 Generalisability / replicability

As Table 16 shows, just over three-quarters of the reviews were classified as assessing the generalisability and replicability of the included evidence (n=24/30; 80%).

In this study generalisability was defined as the ability to apply the results of a health promotion intervention to one's own context (e.g. location; population; health and educational system). Replicability is a broader concept incorporating generalisability, but also the ability to re-create the intervention in one's own context. The systematic reviews were categorised as assessing generalisability if they commented on how the outcomes of the evaluations might apply in other areas, and classified as assessing replicability if they commented on how the interventions might be mounted elsewhere.

Table 17 reports a classification of the different aspects of generalisability and replicability assessed. Most reviews provided a basic description of the characteristics of the interventions and study populations evaluated. Some reviews devoted comparatively more attention to these issues than others, and examples of these are given in the following sub-sections.

Table 17 - What aspects of replicability and generalisability are assessed / extracted (sub-set of 24 reviews classified as assessing generalisability / replicability)

Aspect assessed	**Proportion of reviews N (%)**
Intervention delivery	24 (100)
Intervention content	22 (92)
Characteristics of study population	20 (83)
Infrastructure	10 (42)
Outcome measures	9 (38)

NB. Reviews could be classified as assessing multiple aspects of reliability and generalisability, hence total numbers exceed 30

4.7.1.1 Intervention delivery

The most common aspect of generalisability / replicability assessed was the delivery of the intervention (n=24/24; 100%). Within this category the reviews commonly discussed the setting in which interventions were delivered, and the intervention provider. In terms of setting, a commonly discussed issue was whether or not the same intervention effects would be achieved

if it were delivered in a different location. For example, Kirby (2006: 46), in his review of sex and HIV education programmes in young adults, commented that the interventions reviewed could be considered generalisable to most countries:

"The effects of these programs were quite robust. They were just as likely, if not more likely, to be effective in developing countries as they were to be effective in the U.S. or other developed countries. They were effective in both urban and rural areas, in both low and middle income communities, in both school and community settings"

He was careful to point out that the intervention itself was homogenous and required tailoring to specific contexts:

"Of course, the exact same program was not implemented with all of these groups; rather programs were appropriately designed or tailored for some of these groups"
(Kirby, 2006: 46)

In terms of intervention provider, questions were raised about whether providers would have the same level of skills, motivation and autonomy in practice as in the evaluation setting. For example, Hillsdon *et al* (2005: 8), in their review of interventions to promote physical activity, commented:

"The physicians in the studies based in a primary health care setting may have been more motivated to deliver the interventions than might be observed in a non-trial setting"

A similar issue was raised by Huibers *et al* (2003: 9) who warned that, in practice, providers may lack the skills and expertise of the highly trained practitioners who delivered the interventions in their review:

"These findings should be interpreted with considerable caution: the two studies on PST [Problem Solving Treatment] were conducted by the same research team and groups consisting of only 30 to 40 patients were treated by a small number of experienced and highly trained research GPs, which limits the translation to routine general practice"

4.7.1.2 Intervention content

The content of the intervention was another common aspect of generalisability / replicability assessed, as mentioned by 22 (92%) of the 24 reviews. One issue raised was the applicability of

the intervention to routine practice. Turner *et al* (1996), who reviewed educational and behavioural interventions for back pain in primary care, commented that the kind of interventions they believe would be most applicable in the field had not been evaluated. They discussed in detail the kind of interventions that should be delivered, in terms of provider, length, content, and framework. For example, cognitive behavioural therapy was mentioned as one particularly appropriate type of intervention:

"The studies described in this article suggest that brief educational interventions are not likely to improve significantly the outcomes of people with low back pain seen in primary care settings. Although cognitive behavioural interventions for low back pain would appear to be highly applicable to primary care they have not been implemented or studied in such settings"
(Turner *et al*, 1996: 2854)

Another issue that caused concern was the often poor description of the intervention by studies. Wolitski *et al* (1997), who reviewed the effects of HIV testing and counselling on risk reduction behaviours, found that description of the content of the intervention by the studies was often inadequate.

4.7.1.3 Details of study population

Generalisability / replicability in relation to the intervention recipients was assessed by just over two-thirds of reviews (n=20/24; 83%).

An example of a review that particularly considered this issue was that by Ogilvie *et al* (2004; 2005), who reviewed interventions to promote walking and cycling as an alternative to using cars. They found that the studies included in their review varied considerably, yet provided little detail on the intervention recipients:

"It became clear that both the types of study design and the nature of the study populations varied widely. Some studies had used comparatively robust methods to measure, for example, changes in vehicle flows along certain roads, but these studies could tell us nothing about the people using those vehicles or about their non-vehicular (walking) trips"
(Ogilvie *et al*, 2005: 887)

They consequently devised a two dimensional hierarchy of study quality, taking into account both internal and external validity:

"We categorised studies not only on the study design (a marker of internal validity) but also on the study population, which we took as our primary marker of external validity—in other words, a marker of how useful the study would be for answering our question about changes in population health and health determinants"
(Ogilvie *et al*, 2005: 887)

They stated that the classification system allowed them to better understand how the interventions studied might apply elsewhere. This allowed them to make research recommendations for better quality evaluation in specific sub-sets of people.

Another example was the review of physical activity interventions by Hillsdon *et al* (2005), who made comments similar to those discussed earlier about motivation being higher within the context of a study (Section 4.7.1.1). They suggested that those volunteering to receive the intervention would be more motivated, and therefore more likely to benefit from the intervention, than would be the case outside of the study setting:

4.7.1.4 Infrastructure

Ten (42%) of the reviews commented on the whether or not evaluated interventions could feasibly be replicated in, or results generalised to, other settings. To be classified as such reviews had to have considered the social, economic, and political infrastructure of the settings in which the intervention is intended for use. The reviews tended to devote more attention to economic issues, commenting on intervention costs and resources. For example, Powell *et al* (2004: 4) remarked that the potential benefits of visual acuity screening interventions in school-age children and young people are likely to vary according to the economic status of a country:

"The impact of a screening programme may depend on the economic development of the country in which it is taking place. In more developed economies, where spectacle provision is widely available, the impact may be small. In poorer countries the potential impact may be greater if successful delivery and appropriate intervention can be achieved and maintained"

Hillsdon *et al* (2005: 8), commenting on interventions to promote physical activity, noted that their resource intensive nature may be a barrier to implementation:

"Many interventions provided components which would be difficult to deliver in usual practice as they would demand large resources"

Dunn *et al* (1998), who also reviewed physical activity interventions, specifically aimed to assess whether interventions were cost-effective. They found little data on cost-effectiveness and recommended further analyses to determine the feasibility of physical activity promotion. They speculated on the likely cost-effectiveness of one particular type of intervention from their review, signs in public places encouraging stair rather than escalator or lift use. They suggested it could be a relatively low cost option per person reached:

"The cost-effectiveness of environmental manipulations such as sign use are likely to be one of the most cost-effective interventions that public health agencies could deliver because of the relative low cost of signs and their ability to reach large numbers of individuals"
(Dunn *et al*, 1998: 409)

4.7.2 Explaining results

Reviews were classified as such if they assessed how the outcomes of the included studies varied according to specific characteristics of the intervention and / or the study population. That is, they attempted to unpick the factors associated with the effectiveness (or otherwise) of the health promotion interventions. Any methodology could be used to achieve this, either qualitative or quantitative. As Table 16 reports, 17 (56%) reviews fell into this category.

Explanation of effects was either performed narratively or empirically (Table 18).

Table 18 – Methods used by reviews to explain results (sub-set of 17 reviews classified as 'explaining results')

Method	Proportion of reviews N (%)
Narratively	9 (53)
Empirically	9 (53)
Statistical sub-group analysis	4 (50)
Non-statistical sub-group analyses	2 (25)
Inclusion of process evaluation data	1 (12.5)
Meta-regression	1 (12.5)
Classification	1 (12.5)

NB. One review was classed as using both narrative and empirical methods hence why total numbers exceed 17

Just over half of the reviews were classified as explaining their results narratively (n=9/17; 53%). In these reviews the systematic reviewers proposed theories regarding why the intervention was or was not effective, such as whether effects hinged on particular characteristics of the intervention, the participants, or other factors.

For example, Ciliska *et al* (2000) were interested in identifying differences in effects according to factors relating to the study participants (e.g. age, socio-economic status) and the intervention (e.g. location, theoretical basis). They commented:

"Generally, interventions were most successful if part of a multi-component program, if they included education directed at behavioural change as opposed to acquisition of information, if multiple contacts were made with the participants, and if the message was not generally about nutrition but specifically targeted to the increased intake of fruits and vegetables"
(Ciliska *et al*, 2000: 350)

Just-over half of the reviews were classified as using empirical methods to explain results (n=9/17; 53%). Statistical sub-group analysis was the most commonly reported method of empirical explanation, as used in four of the eight reviews classified (Table 18). This involved assessing how the overall quantitative estimate of the effectiveness of an intervention, as identified using meta-analysis, varied when studies of particular types were grouped together separately. For example, Dishman and Buckworth (1996), in their meta-analysis of interventions to increase physical activity, examined how much the overall effect of the interventions varied according to a large number of 'moderator variables' considered to potentially influence effectiveness. These included participant variables (e.g. age, gender, ethnicity), intervention variables (e.g. setting, provider, length), and type of physical activity (e.g. aerobic; strength; active leisure time). Variables that were statistically significant were then entered into a meta-regression model to assess which of them independently accounted for variation in effect sizes.

4.7.3 Quality of the intervention

Reviews which assessed the quality of the intervention were classified as such if they made any evaluative comments about the development, implementation and characteristics of the interventions included in the review. This is in contrast to comments made about the quality of the evaluation methods, as discussed earlier. As Table 16 reports, just over a third of the reviews were classified on this basis (n=11/30; 37%). Issues raised include intervention implementation, training of intervention providers, ethics and theory.

4.7.3.1 Intervention fidelity

In three of the reviews the quality of the intervention was formally assessed using criteria (Huibers *et al,* 2003; Shults *et al,* 2001; Tingle *et al,* 2003). The criteria covered issues such as whether the experimental and control/comparison interventions were explicitly described; whether there was an acceptable level of 'compliance' with the intervention; and whether the intervention was implemented as planned. In terms of the latter, Tingle *et al* (2003: 65) described the importance of assessing intervention fidelity:

"If a study was not implemented properly or as intended, results from the study cannot be attributed to the intervention. Therefore, if some type of intervention tracking occurred, either through use of observations or tracking forms, the evaluation received a 2. An evaluation that showed a modest degree of implementation tracking received a 1. Evaluations that did not track level of implementation received a 0"

Huibers *et al* (2003: 9) assessed, amongst other things, whether intervention providers had been adequately trained. They noted that reporting of such details was poor:

"Only two studies reported that GPs were supervised throughout the trial...in eight studies it was mentioned that GPs were trained…but only four studies elaborated to some extent on the specific content of the training"

Huibers *et al* (2003: 9) considered that adequate reporting of the intervention is important to enable to readers to understand the results, and judge the validity of the findings:

"This lack of vital information makes it difficult to interpret the results of studies…especially in this field of research, in which the blinding of patients and caregivers is virtually impossible, a thorough description of all factors that might introduce bias is of paramount importance"

Similarly, Stout and Rivara (1989: 377) speculated that failure of interventions may be down to poor quality:

"In the four studies based on surveys there was no control of the content, length, or quality of these programs. The evidence lack of effectiveness may not be due to a lack of efficacy of formal sex education, but rather to the quality of the programs examined"

4.7.3.2 Ethics

On occasion systematic reviewers advised caution with regard to interventions they considered ethically dubious. For example, Dyson *et al* (2005) in their review of interventions to promote uptake of breastfeeding, expressed reservations about the appropriateness of one of the included studies which evaluated a policy of 45 minutes of mother-infant contact immediately after birth followed by complete separation until discharge. They noted that the intervention did not increase or decrease breastfeeding initiation rates and urged caution in the interpretation of findings:

"Generalisation of the result of this evaluation is not recommended due to the moderate quality and size of the study and to fundamental concerns regarding the practice of routine separation of mother and baby prior to hospital discharge. The World Health Organization recommends mothers and infants should not be separated after birth unless there is an unavoidable medical reason"
(Dyson *et al*, 2005: 7)

There were, however, examples where description was considered adequate, and the intervention judged to be of high quality. For example, Rotherham-Borus *et al* (2000: s60), who reviewed the effectiveness of HIV prevention interventions with young people, commented on the merits of one particular study, the US National Institute of Mental Health Multisite HIV trial:

"*Uniquely this project had substantial descriptive information on those recruited, consistently high levels of quality control for the delivery of the intervention and the assessments and very high follow-up rates over 3, 6 and 12 months"*

4.7.3.3 Use of theory

Another issue raised was the theoretical underpinnings of interventions. Twelve of the 30 systematic reviews reported whether or not their included studies were theory-based. Often these theories were used to explain the mechanisms for change in health-related knowledge, attitudes, intentions and behaviour. For example, Hurtsi and Sjoden (1997) reported that of the 24 studies in their review the most commonly used theoretical model was Social Learning Theory. Other models used included the Health Belief Model, and various theories of cognitive learning. Some reviews also examined whether there were differences in effect according the presence or absence of theory, as mentioned earlier (Section 4.7.2). In most of these reviews use of theory as a marker of intervention quality was implicit. However, only two of the 11 studies

classified as assessing intervention quality made explicit reference to theory as a marker of quality (Booth and Watters, 1994; Dishman and Buckworth; 1996). Booth and Watters (1994: 1522) suggested some degree of consensus in the field of HIV/AIDS about the necessity of theory to enable effects to be attributed to the intervention undergoing evaluation:

"Social scientists in general, and AIDS researchers in particular have often emphasized the need for theory in strengthening causal interpretation"

Booth and Watters (1994: 1522) found that the majority of the interventions in their review were not theory-based, limiting the ability to identify the mechanisms contributing to effectiveness:

"Emphasis was placed on whether programs succeeded, not on how they succeeded"

Dishman and Buckworth (1996: 714) affirmed the importance of theory by suggesting that interventions are only 'optimal' when based on relevant theory:

"Another implication of our analysis is that previous interventions for increasing physical activity applied in health care settings, including cognitive behaviour modification, were not implemented optimally. Our qualitative evaluation of the studies suggests this may be explainable because the studies did not use standardised approaches based on newer theories about how health behaviour, specifically physical activity, changes"

4.8 Chapter summary

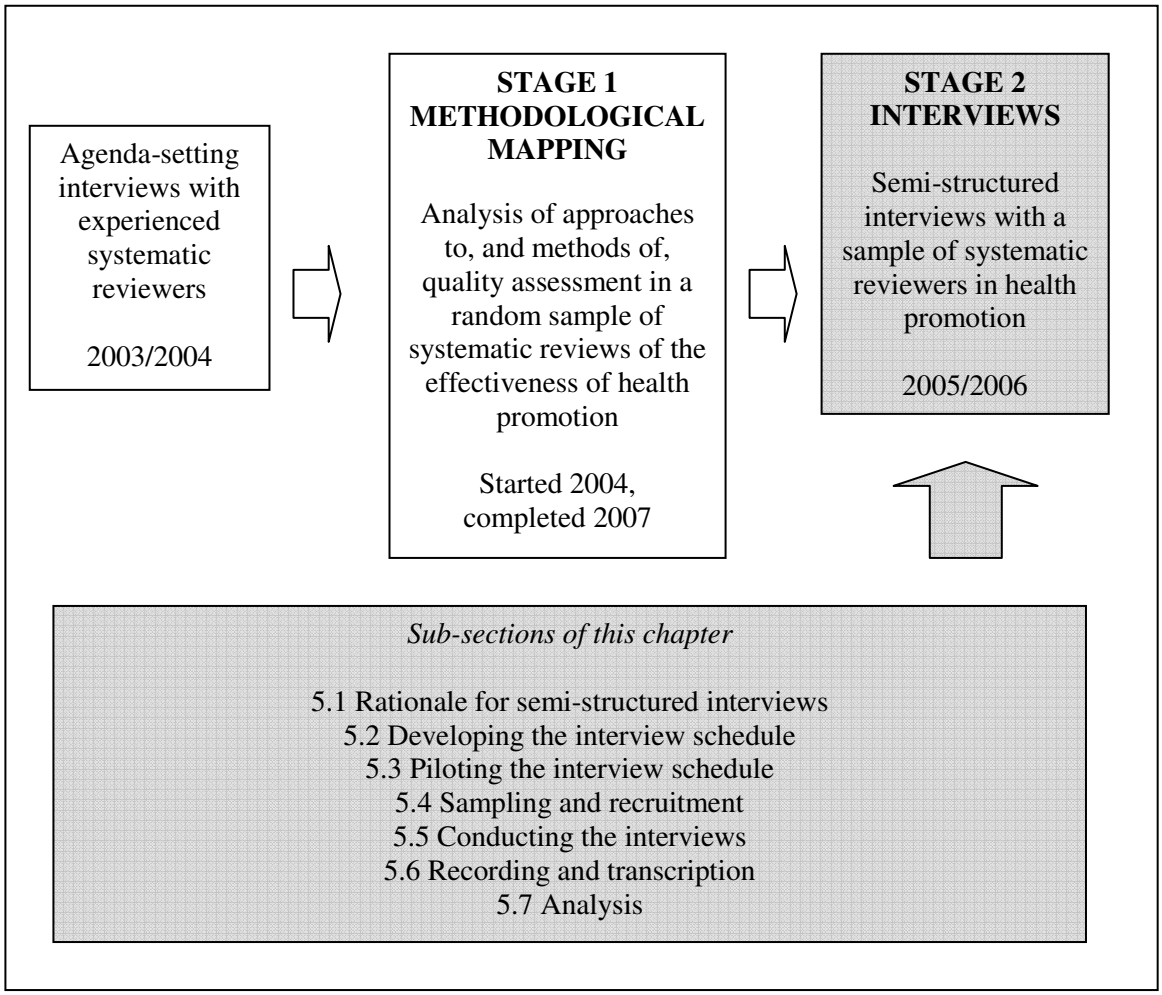
This chapter has presented the results of Stage 1 of this study, the descriptive mapping of a sample of systematic reviews of health promotion. The findings raise a number of interesting issues and have implications for research, policy and practice in health promotion. These are discussed in detail in Chapter 7. The next chapter describes and justifies the methods used for Stage 2 of this research, semi-structured interviews with a sample of systematic reviewers in health promotion.

Chapter 5 - Methods for Stage 2: Semi-structured interviews

Chapter outline

The focus of this chapter is the methods used in Stage 2 of the research, a series of interviews with systematic reviewers. The chapter begins by providing a rationale for using semi-structured interviews, followed by a description of, and justification for, the methods used. These include developing and piloting the interview schedule, sampling and recruitment of interviewees, the process of conducting the interviews, data transcription and analysis (Figure 10).

Figure 10 - Sub-sections of this chapter, and how they relate to Stage 2 and the study in general



5.1 Rationale for semi-structured interviews

There are a variety of ways researchers can elicit information about people's perspectives, experiences and attitudes on a given issue. The approach chosen should be appropriate to the aims and objectives of the research. Interviews are one of a range of data collection methods available to the researcher, alongside questionnaires, focus groups, and observation. Interviews are classified into three main types, (i) structured; (ii) semi-structured, and (iii) unstructured, with an extensive literature on their use (Bowling, 2002; Gorden, 1992; Kvale, 1996; Oppenheim, 1992; Patton, 2002; Rubin and Rubin, 2005). As their names suggest, they vary primarily in terms of their format, particularly the type of questioning employed.

Interviews, as opposed to other methods of data collection, were chosen for Stage 2 of this study for two reasons. First, as mentioned in Chapter 1, some of the research objectives could only be met through interviewing systematic reviewers (research objectives 1, 2, 10 and 11), as the issues in question did not appear to have been discussed in the existing literature. Secondly, systematic reviews are technical in nature and studying their methodology requires flexible methods of data collection. Like many research activities they often involve the use of technical terms, some of which are used inter-changeably. For example, 'selection bias' is commonly defined as a threat to internal validity arising due to differences in socio-demographic and other variables between participants in the experimental and control arms at the start of a RCT (Deeks *et al*, 2003; Grimes and Schulz, 2002). However, it may also be used to describe differences between those who volunteer or consent to take part in a research study, and those who decline. It may further be used to refer to the bias caused by inconsistencies when screening studies for inclusion in a systematic review. In contrast to questionnaires, for example, interviews provide more opportunity for such ambiguities to be clarified (Bowling, 2002). Interviews were therefore particularly suitable for this study.

Semi-structured interviews combine elements of both structured and unstructured interviews. Carter and Henderson (2005: 218) refer to semi-structured interviews as 'planned but flexible'. Bowling (2002: 260) defines semi-structured interviews as containing 'structured questions without response codes'. They are used when the researcher has specific questions which require a greater depth of response than would be achieved using a structured interview, or a questionnaire. Semi-structured interviews, as opposed to structured or unstructured interviews, were chosen for this study, for two reasons.

First, the data generated by semi-structured interviews in Stage 2 were complimentary to those generated from the mapping exercise in Stage 1 of the research. The data extraction instrument

used in the first stage was highly structured, and whilst useful in charting the methods used in published systematic reviews, it was unable to generate the qualitative detail needed to fully meet all of the objectives of this research study. For example, one of the research objectives was to assess who might be the most appropriate agent to assess quality in a systematic review, and how they can be appropriately trained and supported (research objective 9). Yet the published systematic reviews analysed in Stage 1 provided comparatively little detail on the background of the authors, their training, and their experiences of doing systematic reviews. Semi-structured interviews can elicit in-depth information on such issues, through open-ended questions, and careful probing (Bowling, 2002).

Second, it was important to employ some degree of structure to the interviews. Rather than raising general issues for discussion during the interview, as would be customary in unstructured interviewing, it was necessary to develop a sequence of fixed questions covering each of the relevant research objectives. Had the aim of the exercise been more exploratory then an unstructured approach would have been more appropriate.

5.2 Developing the interview schedule

The interview schedule used in this study was devised, tested and refined between August and October 2005 (the final version can be found in Appendix 5). This section describes this process with reference to the methodological literature on research interviewing.

It is crucial that the questions used in an interview are carefully chosen, worded and sequenced to ensure responses that are reliable, valid, and relevant. The type of questions used should be cognisant with the aims and objectives of the research study and should be logical, relevant and motivating (Gorden, 1992). As discussed above, the research objectives for this study required a method of data collection which gives the interviewee the opportunity to talk in detail about their experiences and views, whilst being guided by questions on specific issues. This philosophy underpinned the choice of questions used in the interviews, their wording, and sequencing, as described below.

5.2.1 Types of question

The questions employed in interview studies vary in terms of whether they are broad or narrow, direct or indirect, and open or closed. Choice of question in a study should be governed by the kind of data the interviewer requires, which in turn, should be appropriate to the aims of the study. Open-ended and closed-ended questions both have their advantages and disadvantages. The former is useful in exploratory studies where the aim is to elicit qualitative data in order to

identify themes or explore issues. The disadvantage is that, without intervention from the interviewer, the interviewee may stray into areas that are less relevant to the study. They may also be time consuming to analyse. The latter is suitable where the interviewer is aware of the majority of possible responses, and wishes to note how frequently they are mentioned. Whilst closed-ended questions are easier to analyse, they are not able to elicit responses to the same depth as open-ended questions. It is for this reason that most of the questions used in this study were open-ended, to allow the interviewees to talk at length, thereby providing the qualitative detail necessary to meet the research objectives.

Patton (2002) in his text on qualitative research and evaluation methods, recognises six basic types of interview question. These include basic background questions, questions about people's experiences, their values, their feelings, their knowledge/factual information, and their sensitivity to various stimuli. This study employs four of these, as outlined below.

1) Background questions. The first two questions on the interview schedule were 'warm-up' questions to elicit general information about the interviewee's current and previous roles (e.g. 'Could you tell me a bit about your current job/role?'. The reasons were three-fold. First, it was to generate context within which responses to later questions could be interpreted. Second, it was to help put the interviewee at ease by asking them a general, non-threatening question, as opposed to potentially alienating them by initiating the interview with a more detailed question. Third, it was to help stimulate the interviewee's memory as recall bias is a common problem in research, particularly where the subject of study is historical (Graham *et al*, 2003). Many of the questions asked during the interviews in this study required the interviewee to reflect on their prior experiences, so it was therefore necessary to set the scene as early as possible.

2) Questions about experiences. The majority of questions fell into this category. Their inclusion reflects the strong focus in this study on asking systematic reviewers to describe their experiences of learning and doing reviews. An example is the question 'What have been the biggest challenges you've faced in doing systematic reviews so far?' As noted in Chapter 1, few studies have used these kinds of questions to investigate the methods used to conduct systematic reviews in health promotion.

3) Questions to elicit values. Again, a central concern of this study was to assess the views and opinions of systematic reviewers in health promotion, given the paucity of qualitative data in this area. Examples of these kind of questions included 'What do you see as being the strengths of systematic reviews?' and 'Do you think the training currently available adequately addresses the issues most relevant to health promotion?'

4) Factual questions. The interview schedule also contained questions to elicit factual information about the interviewee. Some of these questions were followed by an evaluative question to allow the interviewee to elaborate further on the issue in hand. For example, Question E3 'Could you describe the instrument / criteria you use to assess quality?' was followed by a question to elicit a value judgement (Question E4) 'Why did you choose this instrument / criteria?'

5.2.2 Wording

Devising effective interview questions requires effort to ensure clarity and accuracy, and to avoid biasing responses. Potential pitfalls include use of emotive language, asking two questions in one, asking leading questions, use of confusing or inappropriate vocabulary, and making assumptions about the interviewee.

Gorden (1992) recommends that language should be as simple as possible to avoid confusion and ambiguity. However, there are occasions when using specialised vocabulary is appropriate, for example, when interviewing a member of one's own professional, ethnic, or religious group. As I was interviewing people in my profession, and given the technical nature of this study, some specialist terms were considered appropriate. That said, it was also important to use them in moderation so as not to make the interview too demanding. Gorden (1992) also suggests that vocabulary used can identify the interviewer as an 'insider' or an 'outsider'. If the interviewee is more likely to feel comfortable talking to an insider then the vocabulary should reflect this. However, a disadvantage is that the interviewee may perceive them to be 'too close', and this may limit what they are prepared to disclose. This is a particular problem if there are political sensitivities within the field. Role relationships within this study are discussed further below (Section 5.5)

5.2.3 Probes

Rubin and Rubin (1995) identify three forms of interview questions used in qualitative studies. These are main questions, probes, and follow-up questions which refer back to earlier responses. The interviews in this study follow this model. The schedule contained a series of main questions, each followed by a probe to be used to elicit further information where necessary. Probes are often used where the initial response to a question has been inadequate or irrelevant, and where the interviewer wishes to clarify and elaborate upon a previous response. Their advantage is that they give the interviewer flexibility to delve deeper into responses as and when they consider it necessary (Gorden, 1992).

Disadvantages include the potential for interviewer bias, particularly verbally (or non-verbally) loading the probe (Gorden, 1992). Over-use of probes may also interrupt the interviewee's train of thought, and increase the likelihood of straying off the topic. Given the broad aims and objectives of this study and the use of open-ended questions, probes were necessary to help guide the interview and ensure relevant responses. A commonly used probe in this study was 'Can you tell me more about this', used to elicit further information following open-ended questions such as 'What factors, in your experience, makes quality assessment easier to do?'

5.2.4 Structure and length

The format of an interview is another important consideration. The sequencing of questions and length of the interview can influence the accuracy and relevance of the responses given. Carter and Henderson (2005) discuss the three stages of a semi-structured interview. The first stage involves introducing the topic and broad aim of the research, negotiating consent and establishing confidentiality, and describing how the data will be used. The second stage involves asking open-ended questions. Questions can progressively become more personal or sensitive if necessary. The third stage involves rounding off, asking the interviewee if there is anything else they would like to add to what they have already said.

An accepted way of starting the interview is to begin with general 'warm-up' questions to put the interviewee at ease and to stimulate their thoughts (Bowling, 2002). Later, more detailed, questions may be more effective once the interviewee has had time to recall previous experiences (Gorden, 1992). The interview schedule for this study was designed accordingly, as discussed earlier (Section 5.2.1).

The schedule contained six sections, each of which dealt with a different theme (Table 19). One of the sections (D) was only applicable to interviewees who had experience of providing training on systematic reviews. There were a total of 33 questions between the sections (27 in the instances where section D was not applicable).

The length of an interview is usually determined by how many questions are asked, and how much the interviewee is prepared to disclose. Naturally interviewees will vary in terms of how talkative they are, as will interviewers in terms of how skilful they are in shaping and sustaining the discussion to an appropriate length. Some interviews may last for several hours, whilst others may be considerably shorter (Oakley, 2000). A balance was needed between eliciting in-depth data, and not demanding too much of the interviewee's time.

Table 19 – The sections of the interview schedule

Section	Title	Number of questions
A	The interviewee's professional background	2
B	General questions about the interviewee's involvement in systematic reviews	10
C	Questions about how the interviewee learned to do systematic reviews, and assess quality	5
D	Questions on the interviewee's experiences and views on providing training on systematic reviews (if applicable)	6
E	Questions about quality assessment	7
F	Questions about the future of systematic reviews / wrap-up	3

The latter was particularly important given that many of the interviews were to take place at a busy conference (see Section 5.4.1)

5.3 Piloting the interview schedule

Piloting is an essential stage of any research project. It can help to identify whether or not the respondent understands the question, that the wording is clear, and is in no way leading. Piloting can also be of practical value, such as helping to gauge the average length of the interview (as discussed), and enabling the interviewer get a feel for how the interview flows.

In this study the primary purpose was to ensure that the interview generated the information necessary to meet its aim and objectives. It was considered that up to five people would be a sufficient number to pilot the interview schedule with. Three people were eventually interviewed in the pilot, all of whom were colleagues of mine who kindly agreed to give their time. Whilst it would have been desirable to have included people with whom I was less familiar, as was the intention for the main study, in the interests of pragmatism it was decided to take the opportunity to involve people who were easily accessible.

Two of the pilotees were research fellows in my department at the University of Southampton, and one was a research fellow at the EPPI-Centre, Institute of Education, London. As well as being easily accessible, all were chosen on the basis that they fitted the criteria for interview. That is, they were a lead or co-author on at least one systematic review of effectiveness (see below for further discussion of criteria for interview). One of them also had experience of providing training on systematic reviewing. They were therefore able to answer questions in

section D of the schedule (see Table 19). It was important to select interviewees from more than one institution as it was anticipated that responses would be influenced by departmental policies, ethos, working practices and culture.

Piloting should continue until the interviewer is satisfied that all necessary modifications have been made to the schedule and that it is suitable for use in the field. In this study minor changes were made to the schedule following piloting. A couple of questions were removed as they appeared to duplicate other questions, and other questions were re-worded slightly.

One of the purposes of piloting of the interview schedule was to assess the average interview length. In this study it was anticipated that interviews would last anything between 30 minutes and one hour. It was noted that the average length of the pilot interviews was approximately 40 minutes, and the interviews themselves were around 45 to 50 minutes on average. The finalised version of the interview schedule can be found in Appendix 5.

5.4 Sampling and recruitment

There are a number of different sampling strategies for selecting interviewees in studies of this kind. These are commonly categorised as probability or non-probability sampling (Bowling, 2002; Davis and Scott, 2007; Green and Thorogood, 2004). Non-probability sampling is used in studies, primarily qualitative, where random sampling is not the objective. For example, purposive sampling is a deliberate non-random method of sampling, which aims to sample a group of people with a particular characteristic. Theoretical sampling involves the generation of conceptual or theoretical categories during the research process, the aim of which is to develop and challenge new hypotheses (Glaser and Strauss, 1967; Strauss and Corbin, 1990). Sampling stops when no new analytical insights are forthcoming. Opportunistic or convenience sampling involves recruiting people who happen to be available to the researcher at a particular time in a given place. The most appropriate type of sample depends on the purpose of the research. Since the nature of this research was exploratory a combination of purposive and opportunistic sampling was used.

Sampling was purposive in the sense that interviewees were sought with experience of producing systematic reviews of health promotion topics. To be eligible for interview a person had to have been a lead author or co-author on at least one systematic review of effectiveness of a health promotion topic. This was necessary because the research objectives of this study specifically focused on the views of people with experience of systematically reviewing within the context of health promotion. Interviewees with experience of systematic reviewing specific

types of health promotion intervention, and different topics were also purposely sought (see Section 5.4.2)

Sampling was also opportunistic in the sense that an international conference on evidence synthesis, the annual Cochrane Colloquium (Cochrane Collaboration, 2005), was used as a means of selecting and recruiting potential interviewees. In early 2005, during the *viva voce* held to assess the transfer of my candidature from MPhil to PhD, I proposed the possibility of attending the conference, to be held in October of that year in Melbourne as a means of accessing interviewees. The examiners considered that this would be advantageous, and strongly encouraged my attendance.

5.4.1 Rationale for conducting the interviews at the Cochrane Colloquium

There were five reasons why the 2005 Cochrane Colloquium in particular was chosen for recruitment, sampling and conduct of interviews, as opposed to the Colloquium on a different year, or an alternative conference altogether. First, it was a means of accessing people aware of the methodological issues associated with conducting systematic reviews, a central focus of this study. Located in a different part of the world each year, the Colloquium is generally considered to be a 'state of the art' forum for evidence-based health. The conference facilitates debate, discussion and reflection on issues relevant to the methodology for producing systematic reviews, and more broadly, evidence and its interface with policy and practice. It is attended by people involved in the Cochrane Collaboration (e.g. systematic reviewers), and a variety of other stakeholders including policy makers, health professionals and health care consumers. Prior to 2005 I had attended the conference in 1997, 1999, 2000, 2001 and 2003, and had presented results of research during some of these years (Shepherd *et al*, 1999a; 2001; 2003). I was therefore highly familiar with the conference programme and the types of delegates commonly attending.

Secondly, it was a way of drawing a sample representative of people from different countries. It was considered important to access a range of views and experiences, particularly systematic reviewers from countries outside of Europe and north America where research cultures may be different (Oakley, 2000). In 2005, 757 people attended the conference, representing 45 countries (as ascertained from the conference delegate list which is supplied at the conference). The majority were from Australia (though predictably, since this was the host country), with a significant contingent from the UK, mainland Europe, USA and Canada. There was also representation from South East Asia, China, South America, and South Africa.

Thirdly, it was necessary to access people with specific experience of systematically reviewing health promotion. As Melbourne is the base of the (former) Cochrane Health Promotion and Public Health Field (now known as the Cochrane Public Health Group), and that representatives from the Field served on the conference organising committee, the Colloquium was to have a distinct health promotion focus during that particular year. The Field, whose role was to support the various Cochrane Review Groups to conduct systematic reviews of health promotion topics, also organised a one-day post-conference workshop entitled 'Cutting Edge Debates in Evidence-Informed Public Health', attended by over 50 people. It was anticipated that both the conference and the workshop would be an unparalleled opportunity to access a large number of international systematic reviewers with experience of health promotion.

Fourthly, I was anticipated that interviewing a number of people in a defined period of time would have advantages. For example, conducting a block of interviews over the course of a week and in one venue would allow me to be more focused and thus potentially a more effective interviewer. There were also potential economies of scale in terms of time saved from not having to travel to meet interviewees individually.

Fifthly, I found that having an 'event' to work toward was motivating in the months leading up to the conference. A number of tasks were to be done, including designing and piloting the interview schedule, and sampling and recruiting interviewees. A deadline helped me to manage my time more effectively. The only drawback was that Stage 1 of the fieldwork, which was in progress, had to be given a lower priority to enable all of the tasks for Stage 2 to be completed on time. It is for this reason that Stage 1 took longer to complete than originally anticipated.

5.4.2 Sampling frame

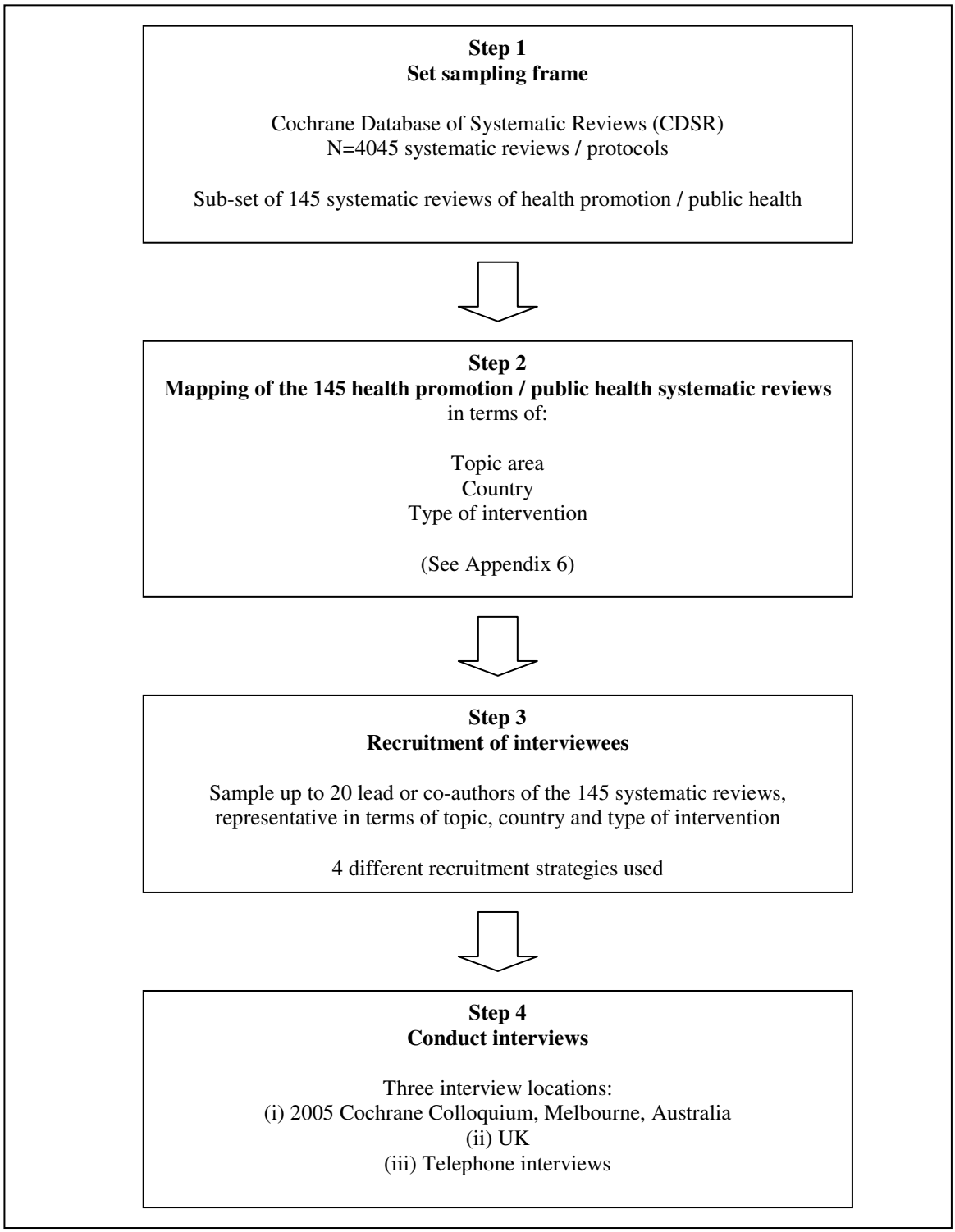
Figure 11 illustrates the key steps followed in the sampling and recruitment of interviewees. This and the following section (Section 5.4.3) describe these steps in more detail.

Davis and Scott (2007: 160) define a sampling frame as:

"A list of units or elements assumed to define best the target population…and it is from this list that the sample is drawn"

Examples of commonly used sampling frames include the electoral register, or list of General Practitioner surgeries.

Figure 11 - Overview of sampling and recruitment



A good sampling frame, Davis and Scott (2007) suggest, is up to date, with distinguishable units, that are counted only once. The electronic Cochrane Library, one of the key resources produced by the Cochrane Collaboration, was the sampling frame in this study (Cochrane Collaboration, 2008; Clarke, 2006). One of the key components of the Library is the Cochrane Database of Systematic Reviews (CDSR) which, as of Issue 3, 2005, contained 4045 systematic

reviews and protocols. Of these, 145 (4%) were classified as being relevant to health promotion and public health by the Cochrane Health Promotion and Public Health Field.

The CDSR, and its sub-set of 145 health promotion and public health reviews, was chosen as the basis for sampling as it is one of the few comprehensive databases of published systematic reviews available. As the 2005 Cochrane Colloquium provided an unparalled opportunity to access systematic reviewers with experience of health promotion, it made sense to use the CDSR and its sub-set of health promotion / public health reviews as a sampling frame.

Alternatives to the CDSR include databases such as DoPHER, the EPPI-Centre's 'Database of Public Health Effectiveness Reviews, described in Chapter 3. Whilst the database was suitable as a means of sampling systematic reviews to be analysed in Stage 1 of this study, the CDSR was more useful to identify interviewees for Stage 2. DoPHER only provides bibliographic details of systematic reviews, as opposed to the CDSR which publishes the entire systematic review itself. Systematic reviews within the CDSR are also more likely to be up to date, as Cochrane policy is that reviews should be updated every two years (Clarke, 2006). This meant that it was more likely that the authors of the reviews could be successfully contacted, and that the conduct of their review would be relatively easier for them to recall. This is important given that recall bias is a particular problem in interview studies (Graham *et al*, 2003). In contrast, contact details of the authors of reviews indexed in DoPHER are those provided in the original publication, be it a journal article, or report. The level of detail is variable meaning it would be harder to trace the original authors.

The downside to using the CDSR to sample interviewees is that they will have been influenced, to varying degrees, by standard policies and procedures for conducting reviews within the Cochrane Collaboration. The Collaboration specifies its recommended methods for conducting and reporting reviews in guidance documents (Higgins and Green, 2008; MacLehose and Remmington, 2005). There is greater potential for a 'Cochrane bias' in the sample than would have existed had the interviewees been recruited via other means. However, despite the Collaboration's desire for uniformity in the production of its reviews, it has been reported that differences in methods do exist between Cochrane reviews (Middleton, 2004; Moja *et al*, 2005). With this in mind it was interesting to see whether any such differences would be reported during the interviews, as this would be suggestive of dissension in methodology, a central concern of this study.

Details of the 145 health promotion and public health reviews were tabulated, to provide an overview of their key characteristics to enable a sample representative of the various topics,

types of intervention and country to be drawn (Appendix 6). The rationale for this was that systematic reviewing methods are likely to vary according to the health promotion topic and type of intervention. For example, more complex interventions involving a range of providers in multiple settings may need to be appraised and analysed by systematic reviewers using different techniques than would be used for more simplistic interventions (West *et al,* 2002). As mentioned earlier, systematic review methods may also vary between countries or regions, reflecting differences in policies and accepted practices of academic institutions and research funders (Oakley, 2000). Having said that, systematic reviews sometimes involve collaborations between individuals in different parts of the world (Doyle *et al*, 2005; McMichael *et al*, 2005), and this may diffuse geographical variations in practice (an issue for this study to explore).

In terms of key characteristics of the reviews:

- The majority were conducted by reviewers in the UK (47%), followed by Australia (14%), the USA (10%) and Canada (8%). Where possible reviewers from a range of countries were sampled, particularly resource-poor and non-Western countries which tend to be under-represented in the field of evidence-based health (although there have been efforts to redress this - Doyle *et al*, 2005; McMichael *et al*, 2005).
- Topics varied from the promotion of cardiovascular health to the promotion of screening uptake, with a strong emphasis on tobacco control/smoking cessation interventions (21%) and injury prevention (15%). The intention was that the sample would represent a range of different topic areas in order to fully explore any variations in systematic review methodology associated with topic.
- The majority of reviews (63%) included studies evaluating more than one type of intervention. Many were multi-component interventions featuring a variety of activities (e.g. education, skills development, service provision), involving a range of providers (e.g. peers, teachers, policy makers) and at different organisational levels (e.g. school, community, region). Again, a sample representative of the range of intervention types, from educational interventions to broader policy interventions, was sought.

5.4.3 Recruitment strategies

To facilitate recruitment it was necessary to identify which authors of the 145 reviews would be attending the conference. For those authors not attending the intention was to take a representative sample of them to interview remotely (e.g. telephone interviews). A challenge, however, lay in identifying who would be attending the conference.

One suggestion was to obtain a list of delegates in advance, and to cross-reference the list with the table of the 145 health promotion / public health reviews to identify which authors had registered for the conference. However, this was problematic since some delegates did not register until just prior to the conference. A further problem was that the conference organisers were not able to release a delegate list in advance due to data protection regulations.

An alternative strategy was to contact the lead or co-authors of the 145 reviews directly to enquire whether or not they were attending. For those that confirmed, a representative sample would be taken and an interview then negotiated. Likewise, for those not attending, telephone interviews could be negotiated with a representative sample. However, approaching each of the authors individually, and the volume of subsequent correspondence this would generate, was considered too time-consuming. To overcome these challenges four different recruitment strategies were implemented, designed to reach as varied a sample as possible and in an efficient manner. These strategies are described further in Appendix 7.

Table 20 provides a breakdown of the number of interviewees interviewed according to each strategy. Just over half of the interviews were conducted at the conference, with the remaining conducted in London (n=5) or by telephone for international interviewees who were not available at the Colloquium (n=2). The recruitment strategy that yielded the most interviewees was snowballing / opportunistic sampling.

Table 20 – Number of people interviewed according to each recruitment strategy

Recruitment strategy	**Number interviewed at conference**	**Number interviewed elsewhere**	**Total**
1. Direct email to representative 10% sample of Cochrane systematic reviewers	2	1	3
2. Email from Cochrane Review Group Co-ordinators	2	2	4
3. Email from the Cochrane Health Promotion and Public Health Field	3	1	4
4. Snowballing and opportunistic sampling	2	4	6
Total	**10**	**7**	**17**

In terms of defining a sample size there is no set number of people to sample in a qualitative interview study. The number of people necessary should be governed by the aims of the study, but also during the study it should be considered whether continued interviewing is likely to yield any further insights (Green and Thorogood, 2004). Kvale (1996) suggests that qualitative social science interview studies commonly tend to include around 15 people, and Green and Thorogood (2004) note that in qualitative interview studies with a specific research question around 20 interviews tend to be sufficient. It was envisaged that up to 20 interviewees would be sufficient to enable this study to meet its research objectives. As the table shows, a total of 17 people were eventually interviewed. Although this study did not employ a theoretical sampling approach (Glaser and Strauss, 1967; Strauss and Corbin, 1990), at around the 15th interview it was considered that saturation of responses was nearing and that further interviews would be unlikely to yield many new findings.

In terms of country, the majority of the reviewers were based in the UK (n=8), followed by Australia (n=4), Canada (n=2), the USA (n=1), South Africa (n=1) and Nigeria (n=1). It would have been desirable to have interviewed a greater proportion of reviewers from resource-poor or non-Western countries. However, there were fewer eligible people from these countries, and for those who were eligible it was not possible to find a mutually convenient time at the conference to conduct the interview.

5.5 Conducting the interviews

Texts on research interviewing stress the value of adequate skill on the part of the interviewer (Bowling, 2002; Gorden, 1992; Green and Thorogood, 2004). This is important for many reasons. The interviewer must be able to recognise when the respondent mentions issues that require further probing, particularly those the interviewer was not anticipating. They must also be able to build up rapport and trust, and to put the interviewee at ease. My training came from the School of Education, University of Southampton whose post-graduate course on Research Methods I attended whilst evaluating peer-led HIV prevention with young gay and bisexual men for my Master of Philosophy (MPhil) degree in the mid 1990s. During the first phase of that project I conducted 50 semi-structured interviews with young men. Further interviews were conducted during the evaluation phase of the project, all of which were analysed and written up in the form of a report to the funder (Shepherd *et al*, 1997a), a journal article (Shepherd *et al*, 1997b), an MPhil thesis (Shepherd, 1998) and a book chapter (Shepherd *et al*, 1999b). During this time I gained valuable experience of research interviewing, particularly as the interviews dealt with the sensitive topic of sexual behaviour. I was able to build on this experience for the

interviews conducted for this thesis. The pilot phase not only helped me to test the questions, but also to reacquaint myself with the practice of interviewing.

5.5.1 Preventing interviewer bias

The methodological literature on research interviewing stresses the importance of maintaining neutrality so as not to influence the responses, either verbally or physically (Bailey, 1994; Bowling, 2002). Methods for reducing the likelihood of 'interviewer bias' include not showing surprise or disapproval of a response, or asking leading questions. Care must be taken over the interviewer's appearance and how much they disclose about themselves and the study. Over-familiarity, for example, may lead the interviewee to make assumptions about the opinions and attitudes of the interviewer, over or under-emphasise particular issues. However, there are circumstances when it is not possible to be entirely neutral, or to maintain a distance from the interviewee. (For example, see Oakley's (1981) seminal reflective account of interviewing socially disadvantaged women).

Given that I was interviewing people in my academic field it was possible that I might influence their responses, consciously or sub-consciously. For example, their knowledge of my academic affiliations or publications could, potentially, prompt them to over-emphasise or avoid certain controversial issues. Knowledge of my academic collaborations (e.g. with the EPPI-Centre) could also potentially be problematic, particularly in the presence of inevitable political sensitivities within the field.

Steps were therefore taken to reduce the likelihood of these biases. Firstly, a decision was made not to submit an abstract to the Colloquium in 2005 to present any of my research. (Although my name did appear as a co-author on a poster presented by a colleague on a topic unrelated to this study (Waugh *et al*, 2005)). Secondly, although the introductory email to interviewees reported my job title and academic affiliation, only relatively brief details were provided about the study. It was important to strike a balance between providing enough information about the study (and not unduly influencing the interviewee), without encouraging them to decline participation. Thirdly, at the start of the interview I provided only a brief précis of my academic background and the research topic. Although at the end of the interview the interviewees were given the opportunity to ask any further questions.

On reflection I consider that there was little likelihood of interviewer bias. I was unfamiliar with the majority of the interviewees prior to the interview (and, to my knowledge, vice versa), and had not formally collaborated with any of them academically. In only one instance did an

interviewee explicitly acknowledge a political sensitivity. With reference to mutual colleagues she felt that they had at times been critical of her work and had made little effort to discuss the possibility of collaborating with her. However, she acknowledged that in reality different academic groups cannot always collaborate on a shared agenda. Although she did not appear to feel the need to withhold her opinion, it is difficult to know whether, had she been unaware of my affiliation with these colleagues, she would have elaborated further.

All interviewees were assured that the interviews were confidential, and responses would be anonymised in the presentation of results.

5.5.2 Location of the interviews

Eight of the ten interviews conducted at the Colloquium were done in a quiet room kindly provided by the organisers. However, as might be expected at a busy conference a couple of the interviews had to be done 'on the hoof' in slightly less suitable locations, either because no rooms were free or to fit in with the interviewee's schedule. For example, one person was interviewed in a quiet part of their hotel lobby, and another on the return flight from Melbourne to London. However, it was not thought that the location compromised the quality of the interview process, as on both occasions there were few other people in the vicinity to interrupt or inhibit the conversation. Of the interviews conducted after the Colloquium, five were done in central London in a quiet room made available to me at the EPPI-Centre, and two were done via telephone.

5.6 Recording and transcription

All interviews, with the exception of one of the telephone interviews, were tape recorded. Each interview was transcribed by a secretary who, once completed, sent me a draft of the transcript to review. In some cases the quality of the recording was impaired due to background noise or momentary dips in the volume of the interviewee's voice. Once a draft transcript was ready I listened again to the interview, partly to re-acquaint myself with what was discussed in order to prepare myself for data analysis, but also to amend the interview transcript with the missing words or phrases as necessary.

5.7 Analysis

A qualitative content analysis approach was used to analyse the data from the interviews. This was chosen as it is a standard and relatively straightforward technique used in the social sciences to categorise and explore qualitative data from interviews (Bowling, 2002; Weber, 1990). Bazeley (2003) describes qualitative content analysis as a method of generating

categories inductively from the material and formulating them as much as possible in terms of that material. This is in contrast to quantitative content analysis in which theoretically pre-specified categories are used. At its simplest, content analysis involves reading interview transcripts and recording the frequency with which key terms and issues are mentioned (Low, 2007). These can be grouped into themes and the relevant dialogue coded according to each theme. It was considered to be an appropriate method for the analysis of the type of data to be elicited in this stage of the research, and was in-keeping with the inductive nature of the study as a whole.

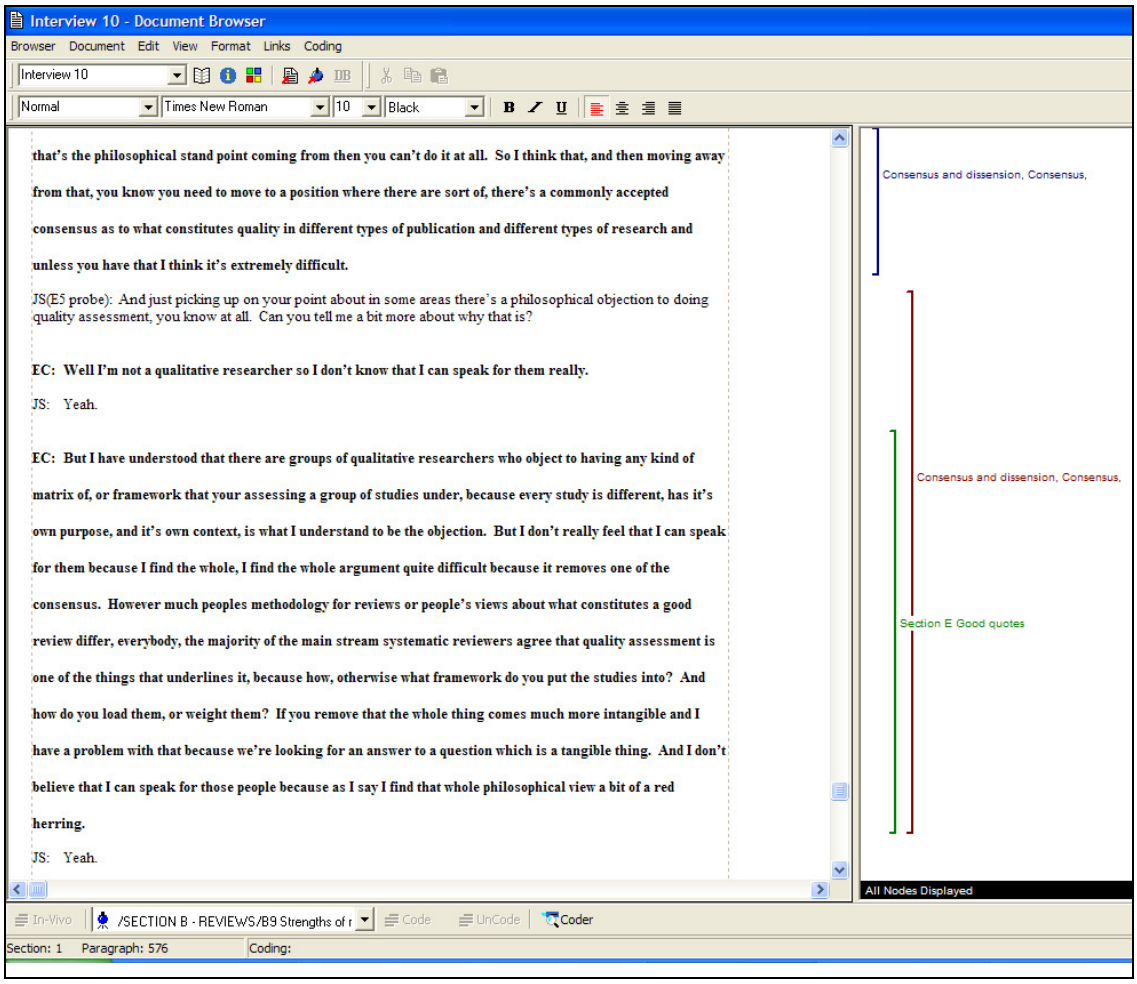
There are a variety of different computer software programmes available for the analysis of qualitative data (Fielding and Lee, 1998; Green and Thorogood, 2004; Weitzman, 2000). The software has made the process of data analysis more efficient, allowing the researcher to store and organise data in a format that is easily retrievable, and which automates techniques that the researcher used to do by hand (Bazeley, 2003). It is important to note, however, that these programmes cannot substitute the considerable skill necessary to identify themes, explore associations and meanings, and to draw conclusions from the data (Green and Thorogood, 2004; Low, 2007). The researcher's analytical and intellectual input therefore remains as important as ever. In this study the software programme NVivo (Version 2.0, QSR international) (incorporating the programme NUD.IST) was used to analyse the interview data (Gibbs, 2002; Richards, 1999). NVivo was chosen in preference to alternatives such as Ethnograph (Qualisresearch) (Seidel and Clarke, 1984) as it was considered comparable in terms of features and capabilities, and also because the University of Southampton supports NVivo with a multi-user licence. This resulted in savings to the project from not having to purchase a copy. I underwent training in using NVivo in May 2006 at a specialist course run by the Computer Assisted Qualitative Data Analysis (CAQDAS) Networking Project, University of Surrey (CAQDAS, 2008).

Each finalised interview transcript was imported into NVivo from Microsoft Word (where it was originally transcribed) and systematically coded in detail. Figure 12 illustrates this process with a section of text from one of the interviews. On the right hand pane are the codes assigned (known as 'nodes', as explained below). As can be seen, a particular passage of dialogue could receive more than one coding.

NVivo has a tree structure that allows the user to construct a series of nodes representing given issues. Nested within each primary node are second-order nodes containing data relevant to the parent node. If necessary, third, fourth, and fifth-order nodes can be added, and so on. In this study nesting as far as third-order nodes was sufficient to categorise the data.

Each of the five sections of the interview schedule was a first-order node (as can be seen on the left hand pane of Figure 13). Each relevant question within each section was a second-order node. The entire interview schedule was therefore accommodated into the tree structure via first and second order nodes. The categorised responses to each question were then assigned as third-order nodes, with relevant quotes from the interviews copied directly from the interview transcripts into these nodes. The software therefore records how many interviewees commented on a particular issue, and permits inspection of what they said.

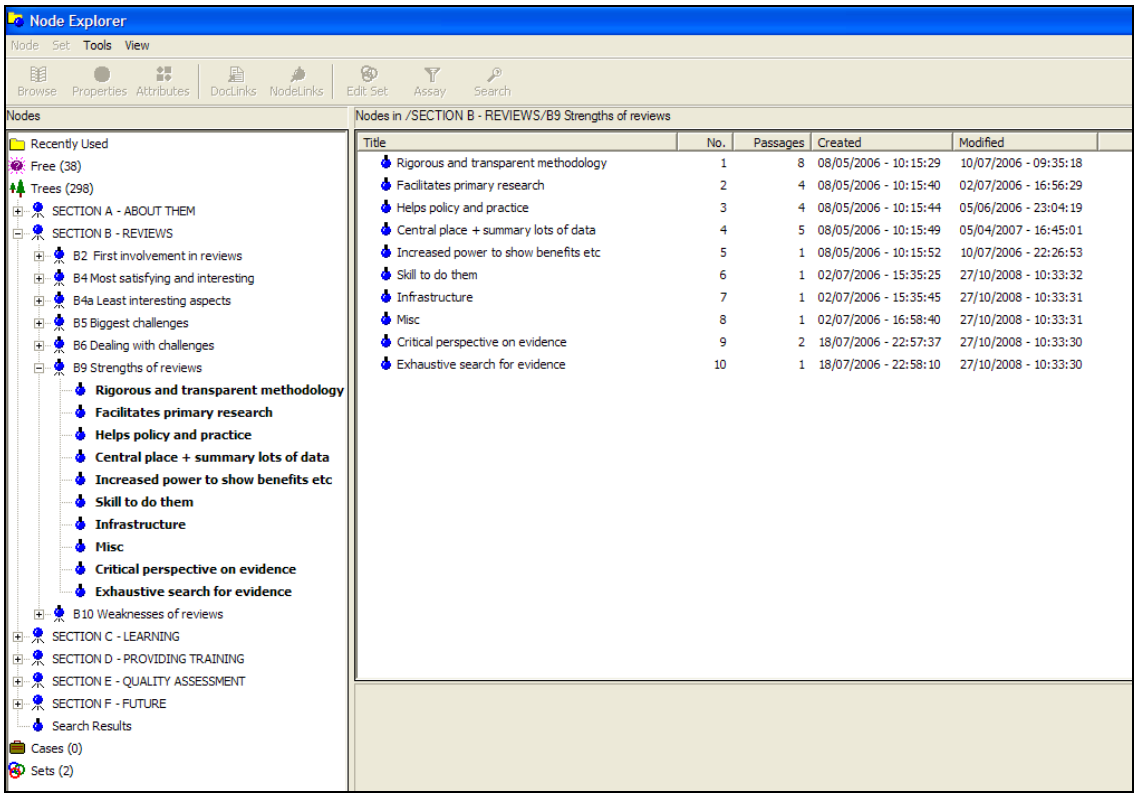
Figure 12 - Example of a coded interview transcript in NVivo



It was the construction of third-order nodes that required most of the analytical input on my part. The categories were constructed through reading and re-reading the responses to each question, identifying emerging categories, and then coding responses according these categories. Figure13 illustrates an example of the categorised responses to the question (and second-order node) 'B9.What do you see as being the strengths of systematic reviews?' Ten different responses were categorised (third-order nodes), which can be seen in the right hand pane. The number of passages of text for each response is displayed in the column 'Passages',

although note that this does not correspond to the total number of interviewees who commented on that particular issue. For example, two interviewees were classified as saying that they thought systematic reviewing 'Helps policy and practice'.

Figure 13 – Illustration of the tree and node structure in NVivo



One passage of text was entered for one interviewee, and three passages were entered for the other. Many of the results tables in Chapter 6 are presented in a similar manner to the tree structure created in NVivo. The example presented in Figure 13 forms the basis of Table 23 in Chapter 6 (Although note that in Table 23 the responses appear slightly different, and are in rank order).

In addition to tree nodes, NVivo also allows the creation of free nodes. These do not permit nesting of sub-nodes, but are designed to accommodate issues and themes that are not relevant to the tree structure. In this study free nodes were created to capture comments made that were not directly relevant to any of the questions asked (and therefore were not coded under the tree structure), similar to what Low (2007) calls 'divergent themes'. As in any qualitative study, it is important to capture and explore unanticipated findings and free nodes allowed these to be documented.

5.8 Chapter summary

This chapter has described and justified the methods used in the second stage of this research, semi-structured interviews with a sample of systematic reviewers in health promotion. It reported the process of developing the interview schedule, and also the sampling and recruiting the 17 interviewees, the majority of whom were interviewed and at the Cochrane Colloquium in 2005. The next chapter presents in detail the findings of these interviews.

Chapter 6 - Findings of Stage 2: Semi-structured interviews

Chapter outline

The chapter begins by briefly describing the characteristics of the interviewees, and their general experiences of, and views on, systematic reviews. It then moves on to describe the approaches the interviewees use for assessing quality of evidence in systematic reviews, focussing specifically on the criteria employed, the rationale for the criteria, factors that facilitate quality assessment, and their suggestions for additional criteria which they consider important for systematic reviews to assess. The final section describes how they learned to do systematic reviews, their experiences of helping others to learn to do systematic reviews, and their views on the adequacy of different methods of teaching and learning.

Interviewees are identified by way of their code numbers in the text (e.g. interviewee 1)

Recap: research objectives relevant to Stage 2

To re-iterate, eight of the 11 research objectives were relevant to this stage of the research:

1. To assess views on the strengths and weaknesses of systematic reviews of health promotion

2. To assess the challenges people have faced when doing systematic reviews of health promotion?

- How have these challenges been dealt with?
- With what success?

3. To assess the extent to which systematic reviews of health promotion assess the quality of included studies:

- What are the barriers to, and facilitators of, quality assessment?

6. To assess the criteria that systematic reviews of health promotion use to assess the quality of included evidence:

- Which criteria are used?
- Why have these criteria been chosen?
- Do these criteria address acknowledged threats to internal validity?

7. To assess whether there is consensus on the criteria by which health promotion evaluations should be assessed in systematic reviews.

9. To assess which types of people commonly participate in the production of systematic reviews of health promotion:

- Who does reviews (e.g. academics, health and other professionals, lay people), and what is their rationale for doing them?
- Who performs quality assessment in systematic reviews? (e.g. people who specialise in producing systematic reviews; people who specialise in the topic area being reviewed; combinations of these)
- To what extent are systematic reviews the product of collaborative teams? What are the advantages and disadvantages of collaborative team working?

10. How do people learn to do systematic reviews of health promotion?

- Which learning strategies are considered most successful?
- What are the barriers, to and facilitators of, learning?
- What are people's experiences of receiving training?

11. What are people's experiences of helping others to learn systematic reviewing?

- What forms of training and support are provided?
- What issues and topics are covered?
- What have been the challenges and successes in providing training and support?

6.1 Characteristics of the interviewees

6.1.1 Academic status

Table 21 - Classification of the academic status of the interviewees (n=17)

Academic status	Interviewees N (%)
Academics	8 (47)
Academic practitioners	5 (29)
Non-academics	4 (24)

Eight of the 17 interviewees were classified as academics (Table 21). Most were employed in universities and ranged in position from junior research fellows to professors. Five were academics who were also professional practitioners (e.g. clinician or other health care professional). Those interviewees not classed as academics (n=4) included three people employed by the Cochrane Collaboration (e.g. review group co-ordination; training officer) and

a hospital based ophthalmologist. Despite not being employed in an academic institution, they all had research experience.

6.1.2 Previous research experience

The interviewees had a variety of research interests and had participated in different types of research. Most had experience of quantitative, rather than qualitative research, and were currently or had previously been involved in clinical, social, or environmental epidemiology (n=8 interviewees). Other forms of research that the interviewees had experience of, aside from systematic reviewing, included primary evaluation of the effectiveness of health interventions (n= 8 interviewees); and surveys and needs assessments (n=3 interviewees). Topic areas that the interviewees had expertise in were numerous, and included HIV/AIDS; accident and injury prevention, obstetrics and perinatal health, and health inequalities amongst others. There was also some interest in research methodology, including methods for producing systematic reviews.

6.2 Routes into systematic reviewing

The interviewees were asked 'Could you tell me a bit about how you first became involved in systematic reviewing?' Table 22 lists a classification of the routes into systematic reviewing mentioned.

A commonly cited route was recruitment to a post specifically to do a systematic review (n=6 interviewees). Of the interviewees who reported this route, three were recruited to posts in the Cochrane Collaboration. Some interviewees also became involved through academic routes (n=4 interviewees) such as PhD supervision. Another commonly mentioned route into systematic reviewing was through expertise in a particular topic area (n=6 interviewees).

Table 22 - Routes into systematic reviewing

Route	Number (%) of interviewees
Recruited into a systematic reviewing job	6 (35)
Expertise in a particular topic area	6 (35)
To advance knowledge in specialist area	2 (18)
'Stumbled' into it	1 (6)
Not reported	2 (18)

Another reason for becoming involved was because the interviewees saw the potential of systematic reviews to advance knowledge in their particular area of expertise (n=2 interviewees). Interviewee 2 commented:

"We know road traffic injuries are a major problem but we wanted to produce some reviews in areas where it seemed that we needed them in order to inform policy makers and also to help them inform the intervention studies…so really became involved because, you know, I saw the usefulness of them as an advocacy tool more than anything"

The interviewees' experience of systematic reviews ranged from those who had participated in a number of reviews, to those whose involvement was relatively recent and limited to just one or two reviews. Three of the interviewees were employed by programmes specifically funded to provide systematic review evidence to inform policy and practice in health promotion, including the EPPI-Centre (London), the Centers for Disease Control (US) Taskforce on Community Preventive Services, and the Effective Public Health Practice Project (Canada).

The interviewees were asked *'Which particular aspects of systematic reviewing have you had experience of?'* All reported that they had experience of conducting all or most of the stages of a review. The interviewees can therefore generally be considered 'all rounders' in terms of their experience of the various aspects of systematic reviewing.

6.3 Perceived strengths of systematic reviews

The interviewees were asked 'What do you see as being the strengths of systematic reviews?' Table 23 lists a classification of the strengths mentioned. The interviewees acknowledged that reviews have many strengths, but most elaborated on one or two as opposed to discussing them all. Whilst some discussed strengths within the context of health promotion, others talked about systematic reviews in general.

6.3.1 Strengths: rigorous and transparent methodology

The strength mentioned most often (n=8; 47%) was the rigorous and transparent nature of the methodology. By this they referred to the ability of reviews to minimise bias in a way that is transparent to all, thereby increasing confidence in the findings.

Some of the interviewees placed a great deal of faith in systematic reviews as a means of delivering a valid answer to a particular question, in particular interviewee 5 who commented that it was the job of reviews to find out 'the truth'

Table 23 - Perceived strengths of systematic reviews

Strengths of systematic reviews	Number (%) of interviewees
Rigorous and transparent methodology	8 (47)
Central repository of evidence	6 (35)
Facilitates primary research	3 (18)
Supports policy and practice	2 (12)
Exhaustive search for the evidence	1 (6)
Increased statistical power to identify significant effects	1 (6)
Critical perspective on the evidence	1 (6)
Cost-effective alternative to further primary evaluation	1 (6)

NB. Interviewees could specify more than one strength, hence total numbers exceed 17

These interviewees therefore believe that one of the key selling points of systematic reviews is that the methodology, which has been developed and refined over time, is based on sound scientific principles. However, a caveat employed by many of the interviewees was that the principles have to be applied correctly, with the acknowledgement that some systematic reviews can be poorly conducted.

6.3.2 Strengths: central repository of evidence

The second most commonly mentioned strength was that systematic reviews draw together all of the relevant evidence in once central place (n=6; 35%). The ability of reviews to assimilate the large volumes of evidence prevalent in certain topic areas was commended, as interviewee 8 said:

"There's so many primary studies that's available, that's been done, all around the world. And I think the big strength of systematic reviews is to bring that together"

Systematic reviews were also praised for making evidence accessible to people unlikely to routinely access and read evaluation studies, such as practitioners:

"From a practitioners' point of view, I always try and think from a practitioners point of view because I've worked with them so much, is that if this provides an overview of an evidence base that they would never otherwise be in touch with" (interviewee 15)

6.3.3 Strengths: systematic reviews facilitating primary research

Three interviewees (18%) were of the opinion that systematic reviews are a vehicle for identifying areas where primary research is lacking. They suggested that it is only by doing a systematic review and charting the evidence base that it becomes clear whether or not there is enough good quality evidence to support the implementation of a particular intervention. Where gaps exist systematic reviews can be a tool for advocating further research. One interviewee commented that this process works best when systematic reviews are commissioned to answer a specific question, as opposed to being commissioned purely because there is a large volume of evidence that has not been assessed:

"You know the gaps that are revealed by asking questions not based on what's out there but what's important. Which is sort of the way you approach it and you don't choose our topics because there's a large literature on it" (interviewee 9)

The issue of whether reviews are conducted to answer a policy question, or purely because there is a body of literature available also arose in relation to perceived weaknesses of systematic reviews, discussed below (see Section 6.4.1).

6.3.4 Strengths: systematic reviews supporting policy and practice

Another strength, mentioned by two interviewees (12%), was that systematic reviews are a useful means of supporting policy and practice. Interviewee 2 considered that reviews help policy makers make better decisions about what is beneficial, as well as what is harmful. Interviewee 9 went further and commented, from her own experience, that the impact of systematic reviews is greater when the review is commissioned within the context of a government health policy programme. She drew a distinction between policy-orientated systematic reviews, such as the government-funded Taskforce that she was involved in, and reviews which are not necessarily conducted for a specific public policy making organisation, such as Cochrane reviews.

6.3.5 Caveats to the strengths of systematic reviews

The interviewees mentioned caveats in relation to two of the strengths discussed above. Firstly although there was acknowledgement of the rigorous and transparent nature of systematic reviews, there were notes of caution regarding the extent to which systematic reviews offer unbiased answers to questions. Interviewee 14 commented:

"They are less biased than many other forms of evidence. I wouldn't say they were unbiased, because I think they are biased, I think bias is inevitable and we really can't get away from, we certainly have to cope with it and be aware of our own biases"

Secondly, despite the view that systematic reviews can facilitate the commissioning of primary research, there was some scepticism about the degree to which they can influence the research agenda. For example, interviewee 12 questioned whether there is an adequate infrastructure to bridge evidence gaps, commenting:

"I think what's a bit disappointing I think from reviews is that there isn't a nice process, how then that gets fed into the primary research agenda. I know its sort of happens in some instances but I think it's more down to luck rather than anything else"

Another interviewee (5) suggested that a there is a lack of incentive for commercial funding of health promotion evaluation, leading to an over-reliance on scarce public funds:

"I think there's a big resistance to randomised studies in health promotion and public health, but the greater resistance is getting hold of the public funds that allow you to do the primary research of high quality…and in the clinic arena you have commercial groups with a stake in the outcome who can help fund the research…and within public health and health promotion we're very dependent on the public purse"

This problem is not unique to the developed world, as interviewee 5, based in Nigeria, commented:

"Another problem is that the randomised trials especially public health trial interventions should be relating to developing countries are not there"

In his view it is in the interests of organisations such as the Cochrane Collaboration to lobby for such evaluation to be funded:

"Because systematic reviews emanate from randomised trials so if there are no randomised trials then there will be no systematic reviews…because I mean it does not make much sense to do a review and then end saying there are no available trials…that makes the question still unanswered, so probably there should be a way to get, to identify trials that have never been done to stimulate or to promote the conduct of those trials for us to have evidence"

There was, however, evidence of efforts to identify research funding, albeit in a developed country. Interviewee 15, employed by the Cochrane Collaboration, reported that she and colleagues considered lobbying a priority:

"And we often see our role is, it's yeah sure it's about reviewing, but it's also about supporting evaluation and particularly just in our local context we try as hard as we can to advocate for rigorous evaluation"

She went on to describe how she and colleagues had established good working relationships with research commissioners and were working towards setting up a multi-disciplinary lobby group.

6.4 Perceived weaknesses of systematic reviews

The interviewees were asked *'What do you see as being the weaknesses of systematic reviews?'* Table 24 lists the categorised responses.

Table 24 - Perceived weaknesses of systematic reviews

Weaknesses of systematic reviews	Number (%) of interviewees
Inappropriate topics + questions	8 (47)
Failure to address deficiencies in primary research	6 (35)
Poor utility	6 (35)
Not user friendly	6 (35)
Negative perceptions of reviews	6 (35)
Time and funding	4 (24)
Deficiencies in methodology	3 (18)
Timeliness	3 (18)
Subjective	1 (6)
Limited impact on the conduct of primary research	1 (6)
Requires great deal of skill to conduct	1 (6)
Tedious to conduct	1 (6)
Difficult to access	1 (6)

NB. Interviewees could mention more than one weakness, hence total numbers exceed 17

A slightly larger number of weaknesses of systematic reviews were mentioned compared to strengths (Table 23). As was the case with the strengths, each interviewee tended to discuss one or two of them in depth. The weaknesses ranged from shortcomings in the methods used to conduct reviews, to limitations in the findings of reviews as a consequence of biases in the primary evaluation studies included.

6.4.1 Weaknesses: inappropriate topics and questions

One of the key weaknesses cited was that systematic reviews do not always address relevant topics. The eight interviewees (47%) who mentioned this had varied takes on this issue. Two remarked that some topics are neglected because reviewers do not perceive there to be enough evidence available to make a systematic review worthwhile. One interviewee (10), picking up on a point discussed above by another interviewee, remarked that areas where there is an abundance of primary research tend to attract systematic reviewers:

"There's a tendency for people to do reviews, especially in certain organisations, people do reviews where they know there's lots of data. So you'll end up with systematic reviews where there are data and then no systematic reviews where they aren't data rather than having some kind of comprehensive overview of what research, what answers we need to what questions first, and then figuring out what research we need to do"

Three of the interviewees remarked that the problem of gaps in the availability of systematic reviews is particularly compounded in public health, where attempts to answer important policy questions concerning interventions addressing the wider determinants of health are seldom made because of a dearth of rigorous primary evaluation studies. Consequently, it was suggested that systematic reviews tend to address questions where RCTs are available, and neglect questions where observational studies are prevalent, partly because of the perception that there will be few studies to include to make a systematic review worthwhile, and those studies that are available will be too complex to appraise and analyse, a point made later in relation to challenges in doing systematic reviews (see Section 6.5.2).

6.4.2 Weaknesses: deficiencies in the primary evaluation evidence base

A weakness reported by six (35%) of the interviewees was deficiencies in the primary evaluation evidence base. Three different issues emerged: (i) absence of primary evaluation evidence; (ii) primary evaluation evidence available but poorly reported; and (iii) primary evaluation evidence available but of poor methodological quality.

Firstly, and as discussed above, in some areas there are few primary evaluation studies available to inform a systematic review. Interviewee 16 commented that, in her perception, one of the reasons users of reviews are dissatisfied with them is because sometimes they contain very little evidence and the review is inconclusive. However, interviewee 12 could see benefits in this situation, suggesting that inconclusive findings about the effectiveness of established interventions should raise important questions about whether it should continue to be used.

Secondly, difficulties in reviewing poorly reported evaluation studies were mentioned by three interviewees. For example, interviewee 1 suggested that reporting bias is one of the biggest problems a systematic reviewer has to deal with. He remarked that a study might selectively report positive outcomes in support of an intervention, neglecting to mention other outcomes that show no benefit or even harm. In his experience it can be difficult to detect where this occurs and it can bias the systematic review. He suggested that in other cases there may be less malevolent reasons for data omission. Restrictions on word allocation in scientific journals often means that certain details have to be omitted. This can limit the ability of the systematic reviewer to fully appraise the methodological quality of a study:

The issue of poor reporting also arose when interviewees were asked to specify any barriers they faced to assessing quality of primary evaluation studies, and is discussed further in Section 6.6.2.1.

Thirdly, interviewee 7 remarked that the poor quality of some primary evaluation studies can have negative implications for systematic reviews:

"To some degree a systematic review is only as good as what it includes… and so it's inherently limited by quality of the included studies which in our experience, especially in the area of injury prevention, it tends to be very poor"

Again, this is not necessarily a weakness of reviews themselves, but if the biases associated with poor primary evaluation studies are not addressed by the review then the weakness is compounded.

6.4.3 Weaknesses: poor utility

Six interviewees (35%) suggested that some systematic reviews lack utility and are not always useful to the people for whom they are intended. One interviewee (16) commented on research

she had done in her area (social policy), specifically examining the role of systematic reviews in decision making. In general she took a cynical view on their contribution:

"The systematic review was all about raising the game of researchers as reviewers and sort of raising a profile of what reviews could offer…and it hasn't come through yet…it's not the case everybody thinks systematic reviews a waste of time. I think, I don't think they've proved their point, utility wise yet"

When asked if she thought this was a widely held view she remarked that she was not sure. However, she was aware of a wider debate about the value of systematic reviews specifically in the field of social policy research and also further afield, where opinions were divided:

"I've always felt that one of the problems with systematic reviews it seems to be the schism between believers and non-believers, that other methods don't necessarily do"

Another interviewee (14), when probed on her views about the extent to which systematic reviews inform decision making, commented that systematic reviews are not the sole consideration in decision making:

"I would say increasingly they do use synthesised evidence and that's a good thing, but it will never be the only thing they bring into consideration because there are so many other pressures on when you make a policy decision…yes I hope that they are increasingly influencing policy but they're never ever going to be the only thing that influences policy"

This comment was echoed by another interviewee (13) who questioned the extent to which policy initiatives are evidence based. He acknowledged that failure to take into account evidence is not necessarily the fault of the systematic review:

"The idea that now that some sort of linear progression between systematic reviewing and what happens in policy is hopelessly off the mark. Many policies were introduced without piloting, even though it would be the systematic review almost invariably say, need for more research, need to pilot, so on and so forth"

6.5 Challenges in doing systematic reviews

A number of issues were raised in response to the question *'What have been the biggest challenges you have faced in doing systematic reviews so far?'* and are summarised in Table 25.

Table 25 - Categorised challenges to doing systematic reviews

Category	Challenge	Number (%) of interviewees
External factors	Lack of resources and time	11 (65)
Analysis and methods	Analysing complex interventions	5 (29)
	Locating the evidence	4 (24)
	Analysing observational studies	3 (18)
	Critical appraisal	3 (18)
	Dealing with heterogeneity and statistics	3 (18)
	Developing the review question	2 (12)
	Lack of flexibility	1 (6)
Evidence limitations	Reporting bias	2 (12)
	Poor evidence	1 (6)
Expertise	Team working	1 (6)
	Lack of specialist knowledge	1 (6)
Misc	Lack of motivation + tedium	1 (6)
	Advisory group	1 (6)
	Interpretation and presentation	1 (6)
	Supporting external groups	1 (6)

Five different types of challenge emerged from the data, which were categorised as: external factors; analysis and methods; limitations of the evidence base; expertise and skills; and lastly, 'miscellaneous' issues not falling into any of these categories. Some interviewees encountered a number of challenges, whilst others just mentioned one or two.

6.5.1 External factors

A lack of time and funds to conduct systematic reviews was a key 'external' challenge (n=11; 65% interviewees). For some interviewees this was their biggest challenge and they elaborated on it at length.

6.5.1.1 Lack of resources

Interviewee 12 described this as a significant issue and reported that he had received little in the way of funding to do systematic reviews. Part of the problem, he noted, stemmed from difficulties in putting together proposals for funding, as well as the perception that there are few funding opportunities:

"I think it's a lack of time, it was probably a lack of time to do the proposals, but that's sort of a generic problem isn't it from any funding. But I do think there's only, I think there's few people who will fund them"

This perception was shared by others. For example, interviewee 11 commented that to be successful in receiving support, one has to tailor the scope of the review to appeal to the funders. In her experience this can be at the expense of doing a review which is truly public health in its outlook, and which is within the research interests of the reviewer. However, she commented that public health is such a broad discipline that it usually can be 'made to fit'.

The perception of one interviewee (14) was that where funding is available it can often be inadequate:

"The Department of Health runs this scheme where they fund people to finish Cochrane Reviews, but it's laughable. They give you £5,000 on, you know what does that buy? Maybe one month of a research assistant"

Opinions such as this emphasise the view amongst many of the interviewees that systematic reviewing is a major undertaking that requires dedicated resources to be done sufficiently. Despite this, three of the interviewees reported that they had managed to undertake systematic reviews without funding, as interviewee 12 commented:

"Nearly all our review work its, in fact all our review work is un-funded. So it's done in, so it's like a hobby I guess. I've never had any funding to do any of my, for any of my involvement in the reviews"

One interviewee (13) suggested that a 'free' systematic review would not necessarily be acceptable in today's academic environment:

"For the first review, which was published in 1997, we just did it gratis and perhaps in a, you know 10 years ago things weren't as goal oriented as they are now"

As already mentioned, some of the interviewees found it hard to seek funds, but even harder to obtain money to enable reviews to be done of some of the more 'complex' types of health promotion intervention. Interviewee 10 commented:

"To really do reviews of complex topics properly would involve much more funding than is usually available and much more time than is usually available"

This suggests that the more complex types of health promotion activity may be particularly disadvantaged in terms of funding for evidence synthesis.

6.5.1.2 Lack of time

Three interviewees talked about their experiences of being given unrealistic time scales for completing reviews. Interviewee 7 commented specifically on the time pressures of doing a public health systematic review, as opposed to a review of a health care topic:

"The difficulty is in researching and obtaining the papers for these reviews, this is why you can do a clinical systematic review, and address a clinical question. We do it in six months, but I don't think we can do a public health review with that shortness of time, because you, the chunk, what takes your time is the searching, obtaining reports and clarification of authors, which takes, in my experience it takes months"

On the other hand, interviewee 12 was critical of systematic reviews which are completed over an excessively long time period, citing the Cochrane Collaboration review process, which he considered to be too protracted. He mentioned receiving peer review comments on his Cochrane review several months after its completion. However, by this time he had changed his job and his new employer was not able to grant him time to revise the review.

6.5.1.3 Solutions

The interviewees were probed to identify how they had dealt with these challenges, and with what success. As mentioned above, one interviewee (11) mentioned strategically tailoring grant applications to make them more appealing to research funders, and reported that this was generally effective. Others mentioned strategies they used to enable them to work more efficiently, such as better time management and organisation of tasks (n=3 interviewees), and working on particular tasks away from the office to reduce the likelihood of interruption (n=1 interviewee). Another strategy mentioned, by interviewee 12, was to enlist the support of others:

"I suppose being realistic about what you can achieve, really relying on people's goodwill, so much goodwill, as there is in a lot of research I guess, but I think particularly with all the reviews that I've been involved with much of it has been down to people's goodwill really. So you're sort of reliant on that, which is, has problems of security, you know I guess"

This opinion echoes many of the earlier views expressed about systematic reviews, namely that few funding opportunities are available, coupled with obstacles to generating successful grant applications, and the fact that many reviews are the product of people's enthusiasm and good-will.

6.5.2 Analysis and methods

Table 25 (above) summarises the challenges categorised under 'analysis and methods' of systematic reviews.

6.5.2.1 Complex interventions

One commonly mentioned challenge was reviewing complex interventions (n=5; 29%). Interviewees found that this presented difficulties, particularly when conducting systematic reviews for the Cochrane Collaboration, which one interviewee (1), felt are more geared towards clinical interventions. She considered it to be a case of 'square pegs in round holes', with Cochrane systematic review methods more appropriate for 'hardcore' clinical topics than 'fuzzier' health promotion interventions. She described potential solutions to this, in terms of new systematic reviewing frameworks, that she and colleagues were in the process of evaluating which 'deal with the fuzzy'.

Also mentioned were difficulties involved in reviewing interventions that aim to instigate changes in the physical environment to promote physical activity (interviewee 14), and policy interventions to promote nutrition (interviewee 9). One interviewee (5) commented on his experience of analysing community based interventions in which cluster randomised trials tend to be used. These had presented particular statistical challenges to him. Fortunately at the time he was on a sabbatical at the UK Cochrane Centre and was able to seek advice from experienced statisticians resident there. In contrast, interviewee 2 mentioned a lack of support which she felt hampered in her efforts:

"I have actually felt a little bit like we're working in a vacuum, in that we are just basically doing what we think is appropriate and with very little guidance from anyone else, even though the EPOC group puts out guidelines they're not necessarily all that useful"

This interviewee had consulted Cochrane guidelines written by the Effective Practice and Organisation of Care (EPOC) group, who routinely systematically review complex interventions. She remarked that she had found it difficult to get advice from people with

relevant experience, and perceived them to be disinterested in the problems she was experiencing.

6.5.2.2 Observational studies

Three interviewees (18%) experienced problems systematically reviewing observational studies. In all cases the topic area being reviewed was such that few RCTs would have been likely to have been conducted. Reviewers had therefore included non-randomised observational studies. They had found it particularly difficult to assess strengths and weaknesses of such studies, as interviewee 2 mentioned:

"We've had a lot of difficulty actually coming up with, you know, very systematic and sort of relatively objective ways of a) extracting the data and b) doing the quality assessment on them, and working out what you can include and what you, you can't"

This was echoed by interviewee 7 who remarked that she did not think there was any consensus within the Cochrane Collaboration on how to assess the methodological quality of observational studies (see Section 6.6.5).

The task of quality assessment in general was mentioned as a particular challenge by three (18%) interviewees. One of them (15) commented that, in her experience, it was mentally taxing:

"I just find it, I mean it's just more, it's a hard thinking task. You've really got to be in the right frame of mind to pick up, well I do, to pick up all the issues, and I guess that's why you have two people doing it"

This is discussed in further detail and within the context of learning to do systematic reviews in Section 6.8.4.2.

6.5.3 Evidence limitations

Another type of challenge was categorised as 'Evidence limitations'. The issues mentioned included deficiencies in the way studies included in systematic reviews are reported (n=2; 12%). Ambiguities in the way data are presented, or omissions of key study details, were obstacles to assessing the strengths and weaknesses of studies, and extraction of results. Consequently, some interviewees had to contact study authors for clarification, and this was time consuming, with time being another key challenge, as mentioned earlier (e.g. see quote by interviewee 7

earlier about the time consuming nature of public health reviews). Poor reporting is discussed in greater detail below in relation to barriers to assessing quality (see Section 6.6.2.1).

6.6 Quality assessment

6.6.1 Barriers to quality assessment

The interviewees were asked *'Do you routinely assess the quality of the studies in your reviews?'* All of the interviewees answered yes to this question. However, three interviewees mentioned that they did not necessarily consider this to have been conducted to a standard to which they would have preferred. In two cases a lack of time was mentioned as a barrier. For example, interviewee 3 mentioned using a sub-set of quality assessment criteria from a larger set in order to save time:

"So we didn't do a, what you would call a 'quality' quality assessment. It was from an EPPI Review that I got the criteria I think. Although I don't think we went as deep as they had gone, we just assessed the outcome studies on 4 criteria, because, you know this proper quality assessment tool for outcome studies is like 4 or 5 pages long, and I've just not got enough time to do that"

Another interviewee (2) felt that a lack of time impeded the standard of quality assessment. With reference to a previous systematic review she commented:

"We could have done it better but I mean again I guess its what we have to say, well I can't actually spare the rest of my life doing a really, really, really good quality assessment because I just haven't got the time, you've got to have, be pragmatic about that"

Poorly reported studies also make the process of quality assessment more time-consuming as interviewee 9 mentioned:

"The papers that are poorly written take 3 or 4 times as long to assess, because you want to be fair, but the information isn't there"

Another interviewee (14) mentioned that problems in accurately deciphering poorly reported studies can result in judgement made at the start of the review being reconsidered towards the end. Again, this increases the time necessary to complete the appraisal.

6.6.2 Facilitators of quality assessment

The interviewees were asked *'What factors, in your experience, makes quality assessment easier to do?'* Table 26 reports a classification of the facilitators mentioned.

Table 26 - Factors that facilitate quality assessment

Facilitating factors	Number (%) of interviewees
Well written-up studies	9 (53)
Good critical appraisal criteria / instrument	9 (53)
RCTs / Experimental studies	3 (18)
Guidelines	2 (12)
Support	2 (12)
Consensus on appropriate criteria	1 (6)

NB. Interviewees could state more than one factor, hence numbers do not add up to 17

6.6.2.1 Well written-up studies

Nine (53%) interviewees commented that having a clear well written study report facilitated the process of quality assessment. Returning to the issue of poorly written studies, interviewee 10 mentioned that they make it harder for systematic reviewers to provide an adequate and fair critique:

"Well primarily I think where you're talking about research the quality of reporting is really important because unless you can get to grips with what the researchers have actually done then you can't assess the quality of it. Things need to be clearly written, their purpose, their aim, their objectives need to be clearly stated, and their methods need to be clearly articulated. It doesn't have to be grandiosely written, in fact it's better if it's not, it needs to be short, punchy and to the point"

However, interviewee 14 conceded that the level of detail required by some systematic reviewers may not have been forefront in the minds of the study authors at the time of publication:

"It's partly about the quality of reporting, yes, and partly because the papers are not written with a view to being included in a systematic review 10 years down the line, you know"

One of the solutions mentioned to the problem of poor reporting involved contacting study authors for supplemental information, to enable a thorough assessment of quality to be made. Three interviewees described varying degrees of success with this strategy. For example, interviewee 5 described success in contacting the author of a study published around 30 years ago:

"The lady was then in South Africa but she has since returned going back to Belgium and she's like 80 something yeah, but she was gracious enough to respond to our email"

Interviewee 1 described a novel approach, whereby study authors are contacted and asked to provide supplementary information. Once the assessment has been completed it is then shared with the author who is given the opportunity to provide any further information, and to comment on its accuracy and fairness. He commented that this has been successful, although it required dedicated resources to achieve, and response rates varied.

Others were not always successful in making contact with authors, such as interviewee 8 who remarked:

"It can be quite frustrating especially around data extraction is when you have missing information, and you're trying to get information from the authors, and there's no response, even after various attempts" (interviewee 8)

One interviewee (11) mentioned that a lack of time meant that she rarely had the opportunity to contact authors at all.

6.6.2.2 Good quality assessment criteria / instrument

Nine (53%) interviewees commented that having access to a reliable instrument facilitated the process of quality assessment.

For example, interviewee 12 suggested that a reliable set of criteria would be one that covers all of the possible issues that might arise during the course of a particular review. He commented on how adjustments needed to be made to the criteria used during the course of one of his reviews to accommodate unanticipated issues:

"I guess that's what happened, particularly in the smoking review we're, you know right up until even now we are near the end of reviewing, we're still making changes to the quality assessment criteria, because things come along and you think, well actually that's not actually what that really means, what we really meant, we want to find out is this"

This was endorsed by interviewee 13 who made the case for planning the criteria in advance, and having a sound justification for doing so:

"I think what's really important is to have a clearly thought through scheme. Be absolutely clear about what it is you're interested in, in terms of bias and so on, and being clear and why you're looking for it. So I think you've got to have a clear intellectual rationale for looking for things, and not to engage in a stamp-collecting episode"

One interviewee (14) remarked that, although having a good quality assessment instrument is helpful, the process often requires an element of personal judgement:

"I mean it is difficult in, you know that there are judgement calls all the time, it isn't you know it, none of these things are nice easy tick boxes and there's so many judgements right the way through"

She felt that using this judgement introduces an element of bias into the methodology:

"Of course it's biased. So it's difficult sometimes to decide, you know whether this is a, you know you had three categories of quality, you know it's often, it really in the end comes down to a judgement call of which category your're going to put it on… I've never met a set of guidelines where they completely dispense with the idea that you make a new judgement"

This was echoed by interviewee 2 who commented:

"I mean if, if there was this great long loop plug of going through for each paper and then you know put it all into a software package and it would spit out an answer it would be fantastic, that's what would make it easier, but obviously there's got to be some degree of subjective"

6.6.3 Types of instrument / criteria used to assess quality

The interviewees were asked *'Could you describe the instrument / criteria you use to assess quality?'* Most mentioned a selection of the key dimensions of methodological quality, rather

than providing an exhaustive list of all of them. Consequently, the frequency with which each dimension was mentioned is not presented here. Table 27 lists the commonly mentioned dimensions.

Table 27 - Dimensions of quality mentioned by interviewees

Dimension
Validity and reliability of the instruments and procedures used to measure outcomes
Sufficiency of the sample size (e.g. whether or not an intention to treat analysis had been performed)
Statistical analysis methods (e.g. whether or not an intention to treat analysis had been performed)
Adequacy of the method of randomisation
Reporting of outcomes (e.g. selective reporting)
Attrition and loss to follow-up
Whether or not selection bias was present
Adequacy of attempts to blind study participants to intervention allocation
Adequacy of the process of concealing intervention allocation

In addition to the above, a number of dimensions of quality were reported that are not necessarily associated with the internal validity of a study, as presented in Table 28.

Table 28 - Dimensions of external validity mentioned by interviewees

Dimension
Whether participants gave informed consent to take part in the study
Integrity of the intervention (i.e. whether the intervention was delivered as planned)
Adequacy of the reporting of the intervention setting
Costs and feasibility of replicating the intervention
Ethics
Duration and intensity of the intervention
Duration of intervention follow-up
Generalisability of the study population

6.6.3.1 Use of existing criteria

Four (24%) interviewees described the quality assessment criteria that they and colleagues had devised. Amongst those who reported using existing instruments, seven specified using instruments associated with the Cochrane Collaboration. In some cases interviewees mentioned using standard criteria recommended in the Cochrane Handbook, focusing on dimensions of quality such as the randomisation procedure, blinding of assessors / researchers, the duration of follow-up, and the intention to treat analysis. Interviewee 10 suggested there was variability within Cochrane in terms of criteria used:

"Different groups within Cochrane use different consensus methods and they don't all concur with the Handbook. So, because I think partly because of different topic areas that have different types of data"

Examples were given of supplemental criteria employed by Cochrane reviewers. Such criteria were used if appropriate to a particular review topic, and / or particular study design. As well as using standard Cochrane criteria, Interviewee 8, for instance, also assessed the setting in which the study was done, whether ethics was approved for the study, and whether study participants had provided informed consent.

The applicability of Cochrane criteria to health promotion was called into question by interviewee 14, who questioned the appropriateness of assessing whether study participants were blinded to the intervention they were assigned:

"Well it's easy if it's a standard Cochrane, because there are standard regulations, you know standard. Except that virtually everything I've done has been in health promotion, and one of the sort of standard ridged Cochrane things is about blinding of the participants to what intervention they're getting. Which of course is completely impossible in a health promotion, you know you can't be blinded to the fact you're being advised to take your exercise, or whatever"

She proceeded to describe how she adapted the criteria to make it more relevant to health promotion. Supplemental questions were included about the length of the intervention (in recognition of the fact that health promotion initiatives can take place over long periods of time), and the setting (in acknowledgement that health promotion can be delivered in a number of different places).

Other instruments / criteria the interviewees mentioned using included the qualitative literature quality assessment criteria by Pope and Mays (interviewee 3), the Jadad instrument for controlled trials (interviewee 12), the Newcastle-Ottawa scale for non-randomised studies (interviewee 4), and the Thomas Canadian tool (interviewee 7). In most cases these were adapted to be applicable to the particular review topic.

6.6.4 Rationale for instrument / criteria used

The interviewees were asked *'Why did you choose this instrument / criteria?'* Table 29 lists a classification of the reasons given.

Table 29 - Justifications for choice of criteria

Reason for choosing criteria	Number (%) of interviewees
Recommended by Cochrane guidelines	4 (24)
Empirical evidence of bias	2 (12)
Background knowledge	2 (12)
Recommended by others	2 (12)
Ease of completion	1 (6)

NB. Not all interviewees reported their justification, hence total numbers do not add up to 17

Four (24%) interviewees mentioned that they used standard Cochrane quality assessment criteria, stating that this had been recommended to them by the Cochrane review group they were members of. Two (12%) interviewees reported that their choice of criteria was based on consideration of empirical evidence of bias. One of these (1) suggested that the findings of empirical studies to investigated bias are mixed and open to different interpretation:

"There's all this stuff about the few things that people succeed in demonstrating have consistent systematic impact on magnitude of effect but nothing's been you know, every time somebody looks at it they see something a little bit different"

Interviewee 13 mentioned that choice of criteria was based on an awareness of bias rather than formal examination of empirical studies:

"Well it was based on, I suppose in a way common sense, in kind of background knowledge, coupled with you know thinking about what the bias is in the particular sorts of studies that we were reviewing"

Two (12%) interviewees reported that recommendations from colleagues or from the methodological literature influenced choice of criteria. For example, interviewee 3 chose to use the sub-set of criteria devised by the EPPI-Centre as, in her perception, this was regarded as having a high status academically. It was also chosen as it was relatively quick to complete in the limited time available to her.

The second interviewee (7) suggested that, in her perception, there is more agreement about the key causes of bias for experimental forms of evaluation (i.e. RCTs) than there is for non-randomised studies. Consequently, it is harder to define what might be the most appropriate quality assessment criteria for such studies:

"So the non-randomised studies, it's just really difficult, we don't have one that we stick to because we can't, there's more fixture about the RCT's we can say that these we know are important, may have been shown to be important"

She and colleagues based their choice of instrument for assessing non-randomised controlled trials from the findings of a published methodological review of quality assessment instruments:

"So for the review I'd just, that I'm working on at the moment, we refer to the Deek's HTA article. We selected one of the tools that they recommended"

This issue was echoed by two other interviewees who recalled the development of their own quality assessment criteria around the mid-1990s. For example, interviewee 17 mentioned that at that time there was little published evidence on what criteria may be appropriate to assess the quality of health promotion evaluations, particularly those which used non-randomised controlled trials:

"It was developed specifically for health promotion evaluations, to look at them, to get away from randomization, 'Yes, No'… and I think it was partly because of there weren't very much, there weren't very many RCTs, so just looking for randomisation wasn't going to be enough"

6.6.4.1 Reviewer's background

Speaking more generally, the background of the systematic reviewer was suggested by two interviewees as being a factor influencing their perspective on which quality assessment criteria are appropriate for health promotion. One of these interviewees (14) suggested that people with

previous experience of evaluating interventions may be more aware of particular study strengths and weaknesses than systematic reviewers who have never done so:

"I mean one thing that changes people's attitudes is whether they've actually done it, been out there and done health promotion, which I never have by the way, being an academic. But you notice that people who come from a very sort of systematic review angle…will be very ridgid about this and so you know. And then people who've come from the, actually 'I've done this'…are looking for different criteria and will ask questions that the very ridged Cochranite wouldn't ask because they didn't, they don't even know that's a problem"

The other interviewee (4) suggested that the varied academic and professional backgrounds of members of her multi-disciplinary team shaped their views on how to judge the quality of studies:

"I'm in a Department of Public Health where I've got people, with a philosophy first degree, and a lot of people with a chemistry first degree, and they're in the one department. I've only got 10 more academics in the department, I've got that kind of diversity. So the chemist, who's an industrial hygienist…you have very different sense of what his criteria, he's got a, he's got an experimental model, and the philosopher has a, a you know a logic, a syllogism, you know of, you know very diverse"

She proceeded to discuss how the introduction of public health policies needs to be underpinned by different types of evidence, from a variety of sources. She acknowledged that it would be difficult for all of this evidence to come from RCTs, and that is why it is difficult to gain consensus on how public health interventions can be effectively evaluated:

"It's big and diverse, that's why it's so much easy just to stick with randomised control trials and be done with it. Or even if you do community trials, cluster randomised, still it's easier. I think once you move beyond it gets very difficult"

6.6.5 Is there a consensus on quality assessment criteria?

The interviewees were asked *'Do you think there is any consensus on what criteria should be used to assess the quality of studies in health promotion?'* Table 30 shows the proportion of interviewees who thought there probably was no consensus, those who felt there was at least some degree of consensus, and those who were unsure.

Table 30 - Interviewees views on whether there is consensus of quality assessment criteria in health promotion

Consensus on criteria	Number (%) of interviewees
Probably not	11 (65)
Unsure	4 (24)
Some degree of consensus	2 (12)

6.6.5.1 No consensus

The majority of interviewees (n=11; 65%) were of the opinion that there was probably no real consensus, and elaborated on why they thought this. Interviewee 10, echoed by many others, made the point that whilst there is some consensus about how one would assess the quality of an RCT, there is less agreement for other evaluation designs:

"Well I think there's no consensus about objective criteria across different types of study or different types of publications, so where you talk about randomised controls there's a reasonable consensus as to what constitutes quality"

She mentioned how some might disagree with the notion that criteria can be used to judge the quality of studies at all:

"I mean in some topic areas, some corners of qualitative research would politically disagree with any kind of quality assessment and say that there are no objective criteria that you can use because all, every different study context etc is different"

However, she made a point of distancing herself from these views, affirming her belief in the need for quality assessment:

"But I don't really feel that I can speak for them because I find the whole, I find the whole argument quite difficult…however much people's methodology for reviews or people's views about what constitutes a good review differ, everybody, the majority of the mainstream systematic reviewers agree that quality assessment is one of the things that underlines it"

Another interviewee (12) suggested that criteria pertinent to health promotion that often get overlooked:

"*I think most people sort of count on one hand, you know the four or five things which they thought were the key ingredients of a trial for example, outside of public health and health promotion. So I think they forget things like context and generalisability which are very applicable to public health and health promotion type studies"*

He suggested that whilst there is core quality assessment criteria applicable to most interventions, there would be supplementary elements applicable to specific types of intervention. Because there are so many different types of intervention that could be classed as health promotion the notion of consensus is highly contextual:

"I think it's fluid, you know because whether you're doing a trial of aspirin or beta-blockers it doesn't really matter. Whereas if you're doing a trial of, oh I don't know, changing the advertising of sandwiches in Tescos, or a trial of referring people to an exercise scheme, they're just two completely different things, so your quality criteria, probably will may need to be very different"

Multiple consensuses

The concept of 'multiple consensuses' was discussed by three interviewees. They suggested that there are a number of different groups of reviewers, who have each developed their own methodology and criteria. For example, interviewee 2 commented:

"I think what happens is that you actually have consensus among a group of people, I mean you have another group of people who have consensus in what they believe but you've got 3 or 4 different camps that believe in different things, so you have to decide where you're going to go and who you're gonna align with"

Another interviewee (14) remarked:

"I think there are various groups working on their own thing and I don't think there's a consensus…I think that if we all got together we'd probably end up having quite a long discussion"

In the view of interviewee 11 there is very little communication between the various groups, and more effort is needed to encourage collaboration in order to reduce duplication and work towards common goals. She cited the Cochrane Health Promotion and Public Health Field as being very helpful mechanism for achieving this. She also felt that one of the groups in question

had, in her opinion, disparaged the work of her own group. Reflecting on the situation she acknowledged that the groups have different histories, and different funders, and so cannot always collaborate on a shared agenda. They may also not necessarily want to collaborate because they want to promote their own work.

6.6.5.2 Some degree of consensus

Only two (17%) interviews suggested that there is some degree of consensus over criteria in health promotion. For example, interviewee 11 commented that many of the quality assessment tools that she had seen were similar, suggesting there must be at least some kind of consensus.

Interviewee 13 commented that the broad nature of public health means that it may be difficult to expect to reach any kind of consensus, citing individual ideas and lateral thinking as being just as important as using formal criteria:

"I mean I'm not sure that that matters that much because people will always do things in different ways. And the idea of forcing an orthodoxy in an area which is incredibly diverse as public health systematic reviews. I mean, I think imagination and lateral thinking bring just as much other help as systematic methods"

Related to this was a comment made by interviewee 14 (in relation to another interview question). She remarked that lay people have different, more subjective views on what constitutes good quality evidence than systematic reviewers (using Cochrane as an example):

"There's more than one way of looking at the world and the, there's the, Cochrane tries to be unemotional, quite rightly, but most of the public evaluate their evidence in a much more emotional way, and is it better? Is it worse? Is it more honest? I don't know, but it's different"

6.6.6 Suggestions for additional quality assessment criteria

The interviewees were asked 'Do you have any suggestions for other issues related to quality that systematic reviews in health promotion should be taking into account?' Table 31 shows a classification of the responses given.

Not all of the suggestions were necessarily about additional criteria. Wider issues were raised relating to the utility of systematic reviews in general.

Table 31 - Suggestions for additional quality assessment issues for systematic reviews to assess

Suggestions for criteria	Number (%) of interviewees
Contextualising interventions	7 (42)
Representativeness / generalisability	3 (18)
Intervention integrity / fidelity	2 (12)
Cost-effectiveness	2 (12)
Compliance / withdrawal from intervention	1 (6)
Broader review questions	1 (6)
Realist synthesis	1 (6)
Qualitative research	1 (6)

NB. Interviewees could make more than one suggestion, hence total numbers exceed 17

6.6.6.1 Contextualising interventions

The most common suggestion for criteria was for systematic reviews to provide more detail on the context within which interventions take place (n= 7; 42%). One interviewee (17) commented on the kind of contextual information that users of systematic reviews find useful:

"And you might not be able to use all these findings because they don't tell you enough about the interventions, or they don't tell you who's involved in them, or anything about what happened when they were implemented and what might have helped, what didn't help"

She also suggested that failure of primary studies to provide adequate details on context inhibits the extent to which systematic reviews can likewise comment such information. This could therefore be a criterion by which studies are judged:

"For reviewers to be able to do them, that kind of review though, they are sort of dependent on what's in the primary studies. I mean you could even link in how the interventions were delivered as a sort of appropriateness indicator"

For interviewee 2 the geographical location in which an intervention is evaluated was a key concern. She remarked that interpretation of the results of a review would be enhanced if more detail was provided on where the interventions were delivered. She gave an example of the use

of motorcycle helmets to prevent head injuries, suggesting how evidence for their effectiveness is likely to vary according to the economic status of the country in which they are evaluated:

"I think they should look at those kind of things and actually try and look at an intervention broadly and sort of make recommendations about how it might fit into a broad, into a broader context. So for example, we can say it's, you know, demonstrable that this, when you have a crash and you're wearing a motorcycle helmet you're gonna have a 70% reduction in head injury, if you were in a high income country wearing a high quality helmet it's not necessarily transferable to someone wearing a low quality helmet you know like in a low income country where you don't have the same emergency services and the same quality, quality helmet and that, that, putting it in that kind of context I think is important in a review to actually point out where, where it's, where it's applicable"

Information on the resources needed to implement an effective intervention was considered to necessary to interviewee 9 who remarked:

"Information on the actual implementation and resources required for implementation sometimes is a little sketchy, and that I think is really important"

6.6.6.2 Representativeness and generalisability

Three (18%) interviewees suggested that systematic reviews should assess how representative and generalisable interventions are to other settings and target groups. The purpose would be to allow users of systematic reviews to gauge whether the interventions could replicated successfully in their own contexts. Interviewee 12 felt that because the relevant data are often not reported by evaluation studies, systematic reviews therefore fail to comment on them. The outcome is that users of reviews are unable to judge whether successful interventions are applicable to them:

"So often, I think generalisability isn't considered by the primary authors and then it doesn't get considered by the reviewers. So there's actually a multiplication of error and then there's the interpretation of the review, in terms of its generalisability for the policy maker" (interviewee 12)

He mentioned that even when the data are reported it may not be clear to the systematic reviewer which characteristics of the intervention and the study population that may or may not be applicable to other settings.

The other two interviewees thought it was important for systematic reviews to assess how study participants had been recruited, and whether this reduced the generalisability of the findings. Interviewee 14, for example, commented specifically about how evaluations, particularly RCTs, often recruit a highly selected group of people, not wholly representative of the target group for whom the intervention might be intended for in routine practice:

"I think a lot of health promotion trials that when you read them very carefully the way they recruit means they've selected a group of people who are sure to succeed, you know and that they're not the great British public or whatever, they're not actually people out there, they're a kind of nice select little group of people"

For this reason she considered it important that systematic reviews should assess and fully evaluate which participants were recruited and which of these actually received the intervention.

6.6.6.3 Intervention integrity / fidelity

Two (12%) interviewees suggested that systematic reviews should assess whether interventions are delivered as originally intended. Interviewee 9, for example, described how this is something that she looks for when appraising studies:

"And then the whole fidelity of implementation, trying to capture that and some studies will say "Yes we had observers" and then ground checks to see how many hours of content on certain topics were provided, and that's always reassuring that there was some attempt to provide the reader with information about, 'Did they really implement the protocol?' and the programme as they said they did"

6.6.6.4 Other suggestions

Table 31 lists a number of other suggestions the interviewees made for issues that should be addressed as part of the quality assessment process. As mentioned earlier, not all of these could necessarily be interpreted as being proposals for criteria to judge studies by. Rather, some were broader suggestions relating to the methods of conducting systematic reviews, their scope, and the interpretation of their findings.

One suggestion, made by interviewee 13, was to broaden the questions posed by systematic reviews. For example, he proposed that reviews should assess the effectiveness of interventions to change policy and legislation, in addition to the more common assessment of the effectiveness of interventions targeted at individuals.

This interviewee alluded to the fact that, whilst these kind of reviews would have a higher profile politically, they would present greater challenges to systematic reviewers on account of the fact that the evidence base would be poorer, requiring a different conception of quality:

"One has to be prepared to sacrifice methodological precision and exactitude for much more arm waving types of reviews. Which can nonetheless be systematic but in a rather different way"

This issue was echoed by interviewee 16 who, discussing one of the tenets of the concept of 'realist synthesis', commented on the belief that the utility of a study is relative to its quality. That is, the topics that may be of particular relevance to policy and practice are often those for which the methodological quality of the evidence is the poorest.

"So I suppose for me this is about the terrible dilemma there is in research, which is that some of the most interesting pieces of work, the most useful are of poorer quality"

Speculating on the reasons for this she questioned whether it is due to interests of the evaluators:

"Was it to do with the sort of people who do relevant work? Are they less interested in doing research well?"

Two interviewees commented on the need for systematic reviews to report information on cost-effectiveness, although it was acknowledged by interviewee 15 that relevant data are infrequently reported:

"But our other thing that we would like to be beefing on about is economic costs, economic analysis, some kinds of statement in reviews around cost-effectiveness of interventions. Obviously this is hard because most of the papers don't talk about it, but I think this is something that we really need to be pushing for"

6.6.7 Who is involved in producing systematic reviews, and in quality assessment?

The interviewees were asked *'Can you tell me a bit about how quality is assessed?'* This was an introductory question within the quality assessment sub-section of the interview schedule, designed to encourage the interviewees to talk in general terms before more specific questions were asked. Five (71%) of the interviewees described who was involved in their systematic

reviews in response to this question (amongst other responses). A further two discussed participation in reviews, but in relation to other questions. The following sub-sections describe participation in systematic reviews mentioned by these seven interviewees.

6.6.7.1 Self-conducted reviews

Three (18%) of the interviewees described systematic reviews that they had done largely by themselves sometimes with additional input from a second person. For example, interviewee 12 remarked:

"Whereas other reviews that I've done, the pressure stocking review was me and a student… it was a much small-scale thing"

Interviewee 3 described the first review that she had done, which was part of a larger evaluation of a men's health initiative. Although the review had been planned by the project investigators she was recruited to conduct it, and did so largely without assistance although under supervision.

Interviewee 5, based in Africa, discussed how he worked largely alone because of geographical isolation. He described how certain aspects of his systematic review, such as quality assessment, were conducted via email with a colleague in Australia:

"I was in Nigeria, the other was in Melbourne so we did it, it independently and then we compared to see if there is any difference between our assessment…we wrote emails and then finally agreed so it was done independently"

6.6.7.2 Team working

Systematic reviews were also said to be conducted by larger teams of people. For example, interviewee 12 reported that his team comprised four people, whilst interviewee 9 mentioned that hers included seven.

Six (46%) of the interviewees mentioned that their teams had been multi-disciplinary. In general the teams comprised people from a variety of backgrounds, all of whom contributed expertise either in the topic under review, or in the methods of systematic reviewing. For example, interviewee 12 described the members of the team he had been involved in who produced a systematic review on the relationship between smoking and eye disease. The team comprised people with academic, health professional, information science and statistical expertise.

A distinction was made between teams such as this, which were formed specifically to conduct a review of a particular topic, and pre-existing teams of people who routinely systematically review a range of different topics. In terms of the latter, interviewee 17 commented on the multi-disciplinary background of members of her pre-existing team (a team which I had previously been a member of). She mentioned that the team routinely conduct systematically reviews of diverse topics, which members are not necessarily that familiar with, and that this can sometimes be a challenge:

"It's like understanding the terminology and the concepts that people are using. But then if you've got the team you can get, it doesn't seem to matter so much, if you've got a team that's got a variety of, cause there's always different perspectives anyway on these things"

She commented that the complimentary backgrounds of the team members often helped overcome these problems:

"So for example working with [name removed] who's got a psychological background. You've got a geography background haven't you? And me with a biologist background, well a perfect combination"

The interviewee also felt that, despite the benefits associated with the team members having different backgrounds, there remained a need for input from experts in the particular topic area:

She gave an example of a systematic review of sexual health promotion interventions for men who have sex with men (MSM) that the team (including myself) had conducted. She considered that the review had benefited from the input from an advisory group of experts in the field:

"I thought that worked really well, but they were gentle with us, I think maybe it was partly that group. But they had some ideas about what we should be doing, and they pointed us in certain directions and towards certain literatures by some of the terms they were using"

However, she felt that one of the downsides of involving expert advisers was that some people might interpret the fact that the systematic reviewers themselves are not experts in the topic being reviewed as being a weakness, undermining their credibility.

6.6.7.3 Assigning team roles

The way in which tasks were distributed amongst team members was described by two of the interviewees. For example, interviewee 12 reported that most members of the review team have an involvement in key tasks such as data extraction and quality assessment:

Rather than nominating particular team members to undertake specific tasks such as quality assessment, the interviewee reported a more informal arrangement whereby people volunteer to participate in whatever tasks they feel able to:

"Yeah, people put their hands up to do it. I think its, it comes down to this willingness of being able to find people who will help you, rather than being in the situation to,'Oh you're qualified we'll let you help"

However, when probed the interviewee expressed what attributes he would look for in someone who would be assessing the quality of studies:

"But I guess in terms of who you might approach, I guess in my head you would, I would have an implicit assumption about the potential skills of someone or not, so for example, you know I wouldn't ask a under-graduate student, for example, because they haven't got those skills"

6.6.7.4 Advantages of team working

In describing the ways in which their teams produce systematic reviews the interviewees commented on what they considered to be the benefits of team working. Interviewee 12 particularly enjoyed working with people from other disciplines:

"I particularly like multi-disciplinary working. So I find that really interesting as well, the different perception or focus that different people bring to a review team is quite interesting"

As mentioned above in Section 6.6.7.2, the multi-disciplinary nature of some teams also helped reviewers come to terms with the technicalities of topics they were not familiar with, a point made by interviewee 17.

6.6.7.5 Disadvantages of team working

Some of the downsides of team working were mentioned. Interviewee 14 suggested that successful team working takes practice and requires open discussion, particularly when undertaking tasks such as quality assessment that require shared agreement.

"I suppose one thing I've learnt off reviewers is pick the people you do the review with very carefully [laughs]…I mean if you're trying to do a systematic review with somebody who doesn't have that concept of academic debate and agreement then you're sunk…I mean if they're sort of prima donnas who, you know 'it has to be right because I said it was right' then you don't get anywhere"

Another interviewee (12) commented on the challenges she had experienced trying to reconcile the diverse interests of the members of her multi-disciplinary review team:

"I think one of the biggest challenges is working with a team that is multi-disciplinary on very broad public health topics, like improving the, you know, the nutritional behaviours of the US population with respect to like fruit and vegetable intake, and when you have a, you know a economist, a psychologist with very different ideas about levels of intervention, or different, frankly different interests…with different professional interests, to bring that group around to agreeing on a research topic for the systematic review and then sort of conceptualising it" (interviewee 12)

One of the strengths of team working was also considered as a possible weakness by interviewee 11. She mentioned how she and colleagues had worked together on a number of systematic reviews over the years, and had reached the stage where they had a high level of inter-rater reliability when independently assessing the quality of studies. Her worry was that because they think in similar ways, certain strengths or weaknesses of studies may go unnoticed. To reduce the likelihood of this happening each systematic review team is joined by an external person, usually a public health specialist taking part in the training scheme her department runs. Both the external reviewer and the established team member would independently critically appraise the same studies, and compare judgement.

Interviewee 17, in contrast, reported that inter-rater reliability in her team was not always high, due to differences in the interpretation of their quality assessment criteria:

"I think, well my experience of talking to other reviewers here is we've all got a slightly different understandings of our own criteria so…so it would be good if we had more in-house consensus work on what the questions really meant, and what was the most important thing to look for in a question?"

6.7 Learning to do systematic reviews

The interviewees were asked *'How did you learn to do systematic reviews?* Table 32 presents a classification of the various ways the interviewees learned to do systematic reviews.

Table 32 - How the interviewees learned to do systematic reviews

Method of learning	Number (%) of interviewees
Practice	11 (65)
From colleagues and mentors	11 (65)
Training courses	10 (59)
Literature and written resources	9 (53)
Applied existing research skills	5 (29)
Academic course	3 (18)
Supervising and teaching others	2 (12)

NB. Some interviewees reported more than one method of learning, hence total numbers exceed 17

Learning through practice, support from colleagues and training courses were commonly cited ways of learning, often in combination.

6.7.1 Learning through practice

Eleven interviewees (65%) specified hands-on experience of doing systematic reviews as a method of learning. Often this was the main method of learning, as interviewee 12 commented:

"Self taught, really. I don't think I've ever been on a course, haven't ever been on a course for systematic reviews. I guess during my Masters in Philosophy I did some modules at the University in epidemiology and I suspect the one on RCTs would have included a session on meta-analysis or systematic reviews. But 99% of it has been on the job training and self-learning"

When probed about how useful they considered learning through practice to be, the interviewees considered that it was particularly beneficial, and that other methods such as reading about systematic review methodology had limitations. For example, interviewee 9 remarked:

"I don't know how else you could explain, I mean you can certainly read about reviews, until you do a review… You don't know what it's like. I mean you don't know what you're getting into, I mean I've seen a lot of discouraged people thinking well I'm going to do a systematic review on XYZ and they have no idea what it involves"

6.7.2 The role of colleagues and mentors

Eleven (65%) interviewees mentioned that colleagues and mentors had helped them to learn, and in the majority of cases this was considered beneficial. They tended to learn from working with more experienced systematic reviewers. For example, interviewee 7 worked with the Co-ordinating Editor of the Cochrane review group she was associated with to produce her first review:

"When I did my first review I had, I worked very closely with our Co-Ed, so although I was leading the review it was more of a joint thing really because this was my first one"

Learning also came from more experienced reviewers within a team. For example, interviewee 9 described how she had been invited to join a team to do a systematic review which helped her to take more of a lead role in a subsequent review:

"I was invited to join a team, led by an experienced reviewer…and so I was brought in on this person's team for a review and then kind of learned, and then when I was in charge of my first review this individual provided guidance"

Learning from peers was also mentioned. Two interviewees remarked that they began systematic reviewing at a time when there were few training opportunities available and self-teaching was necessary. Interviewee 11 described this as being 'the blind leading the blind', and told how she and two colleagues completed their first systematic review (in 1996) largely on a 'trial and error' basis. The other interviewee (15) made reference to learning by 'muddling along' (interviewee 15).

Although opinions on mentorship were generally positive not all interviewees considered that it was beneficial for them, such as interviewee 12:

"I think if somebody's doing a review for the first time they really need mentorship. I suppose maybe that's what Cochrane tries to develop, because they try and link you in an experienced reviewer, I'm not sure how that works, it hasn't worked well for me, it just depends"

6.7.3 Training courses

Ten (59%) of the interviewees reported receiving training in systematic reviews. The extent to which they participated in training varied. Some had only attended one or two brief sessions, whilst others had participated in longer, more detailed, courses.

Five of the ten interviewees (50%) reported receiving training provided by the Cochrane Collaboration. These included various short courses on specific aspects of systematic reviews (e.g. how to write a research protocol) as well as workshops held at Cochrane Colloquia. Various other non-Cochrane training courses were mentioned. The courses attended covered a range of topics, spanning many of the stages of a systematic review. Three (30%) interviewees mentioned attending in-house training sessions in their workplace. In one instance the training was peer-delivered (interviewee 15).

Another interviewee (17) mentioned that an outside expert in statistics had been brought in to their department to train them to do meta-analysis within the context of health promotion. Some interviewees mentioned that it was only after participating in systematic reviews that they received training (e.g. interviewee 13).

6.7.3.1 Access to training

None of the interviewees reported major problems in accessing training, and tended to make use of training courses whenever they became available, particularly if they were being held locally. For example, interviewee 2, based in Australia, remarked:

"I've been to a couple of Cochrane Australasian Centre courses, and, and one I, there was an Australasian Cochrane Conference in Sydney last year which I went to… I go to local methods workshops whenever they're available"

Interviewee 5, resident in Nigeria took advantage of training whilst on a sabbatical in Oxford, funded by a scholarship. He questioned whether he would have been able to access this training if he had not received the scholarship:

"I was wondering if I, if I didn't get an opportunity to be in Oxford how would I have been able to do these things?"

In the opinion of interviewee 12 training opportunities were adequate, although not everyone may be able to afford them:

"I think there's plenty of opportunities for training in terms of courses. Firstly I think there's quite a, you know I see lots of courses advertised on systematic reviews, whether people can then get the time or the funding together is a different issue I guess"

6.7.3.2 Adequacy of training

The majority of comments on the training received were positive. For example, interviewee 16 noted:

"I always find training really helpful, I mean it's never, it's often not immediately obvious how it's helping you but you know you can get something out of what I've been doing in the long run"

When asked if he thought that training adequately prepared her for doing systematic reviews interviewee 14 remarked that it could not be a substitute for practical experience:

"I don't think it ever adequately prepared before you start doing something because you learn so much from doing"

Interviewee 12 remarked that training was not necessarily the best method of learning for him:

"I've never really felt the need to actually go on them, I've looked into the courses, but I've spoken to people like [name of colleague] *about the value and sort of, I guess some people learn well and then there are other people who need to go on a course to get them, given to them, I guess"*

Only interviewee 4 mentioned training as being unhelpful, but this was primarily because it was an introductory course and her knowledge was more advanced.

6.7.3.3 Additional benefits from training

Training had other benefits in addition to the knowledge and skills gained, as interviewee 2 commented:

"I mean but they're good to meet other people to do the networking and to get some of the practical tips about, about, about various things"

6.7.4 Literature and written resources

Nine (53%) interviewees mentioned using literature and written resources in their learning. Often they described augmenting the knowledge and skills they acquired from other sources (e.g. colleagues, training, practice) with information from text books and guidelines on systematic reviews. For example, the Cochrane Handbook (Higgins and Green, 2008) was mentioned by five (56%) interviewees. Other texts cited included the 2001 text book on meta-analysis and evidence synthesis by Mathias Egger and colleagues (one interviewee) (Egger, *et al*, 2001), the EPPI-Centre guidelines (one interviewee), a series of guides published in the British Medical Journal (one interviewee) (e.g. Greenhalgh, 1997), and the University of York CRD guide for carrying out systematic reviews (one interviewee) (Centre for Reviews and Dissemination, 2009). Interviewee 3 commented on the Cochrane Health Promotion and Public Health Field's guidelines as being very useful:

"Well I did get the, the Cochrane, not so much the main handbook but their Health Promotion and Public Health guidelines…I think the Cochrane handbook is as comprehensive as, as any thing"

In contrast, interviewee 12 suggested there was a lack of adequate guidelines on producing systematic reviews of health promotion and public health, particularly about how to do critical appraisal and narrative synthesis.

The other form of written resource used was training materials specifically designed for individual learning, mentioned by one interviewee:

'Most of the, the in-depth training happened when I got involved with the South Africa Cochrane Centre. Where I went through the training material and the training resources that's available on the Cochrane Collaboration Website' (interviewee 8)

6.8 Helping others to learn systematic reviewing

Thirteen of the 17 interviewees (76%) reported that they had provided some form of training and support on doing systematic reviews to others. The proportion of their time spent doing this varied. In four cases it was, or had been, their full time role. In all other cases it had been only one aspect of their work, alongside other activities such as doing systematic reviews.

6.8.1 Types of training and support provided

The interviewees were asked *'Can you tell me a bit about the training you provide?* Table 33 shows a classification of the forms of training and support mentioned.

Table 33 - Type of training and support provided (sub-set of 13 interviewees who provided training)

Type of training and support	Number (%) of interviewees
Professional training	9 (69)
Academic degree course	8 (62)
Mentoring	3 (23)
Cochrane training	3 (23)
Journal clubs	1 (8)

NB. Some interviewees provided more than one type of training and support, hence total numbers exceed 13

6.8.1.1 Professional training

Nine (53%) of the interviewees mentioned that they had taught systematic reviewing to professionals. In general the trainees were health professionals, including doctors, nurses, nutritionists, physiotherapists, health service managers, and policy makers. One interviewee (10) also reported providing training to social workers, and another (17) to professionals working in education.

6.8.1.2 Academic training

Eight (47%) interviewees had taught systematic reviews as part of an academic degree. In most cases the interviewees taught post-graduate students studying subjects such as epidemiology, or public health. However, some taught at under-graduate level to nursing and medical students, and students studying health sciences.

In most cases systematic reviews and evidence based health were reported to be only one component of the syllabus. The detail to which the training could cover these topics was therefore considered to be limited. However, in at least two cases interviewees described teaching a whole course dedicated to systematic reviewing.

6.8.1.3 Mentoring

Three (18%) interviewees reported participation in mentoring programmes. For example, interviewee 10 mentioned her involvement in a programme funded by a benefactor which allows a systematic reviewer from a developing country to spend three months visiting the UK Cochrane Centre in Oxford learning how to do systematic reviews. Her role was to provide one-to-one support to the trainee. Interviewee 15 mentioned a scheme, at the time in its infancy, whereby a novice systematic reviewer is paired with a more experienced mentor within their geographical region. The scheme focuses in particular on helping to teach systematic reviewing skills to individuals in the developing world. Her role was to identify potential mentors and negotiate their involvement. The third interviewee (8) mentioned the 'Reviews for Africa' programme in which novice reviewers in African countries are paired with an experienced Cochrane systematic reviewer.

6.8.2 Content and format of training provided

The interviewees were asked '*Which aspects of systematic reviewing does your training cover?'* The training tended to cover most of the stages of the production of a systematic review, with variations in terms of length and level of detail. The shortest courses tended to last around a day, and covered the principles of evidence-based health and key stages of a systematic review. Interviewee 16 described this as 'entry level' training, designed to enable participants to know where to go for further more detailed training if necessary. At a more intermediate length was the kind of training described by interviewee 17:

"We do a three day course, the first day is a whiz way through all the stages in systematic reviews and the purpose for them and principles and who involved and that kind of thing. And then the second and third days are on the detail of actually what actually needs doing in a protocol and working through a question of people's own choice, how they might search and then jumping on databases and searching and screening and a bit on synthesis and a bit quality appraisal"

However, the extent to which this training went into detail on aspects of systematic reviewing such as quality assessment was reported to be limited, as discussed below. (see Section 6.8.3)

At the other end of the scale were more lengthy courses such as the Masters in Public Health, described by interviewee 1, which featured a whole module on evidence-based health and systematic reviewing. The course educates students on each stage of a systematic review, from start to finish, and requires them to undertake a systematic review as part of the course assessment. The philosophy behind the format was that practice-based learning is an effective way to learn, as the interviewee commented:

"And from the beginning we said the way to learn to do a, way to learn about systematic review is to do a systematic review"

Another example of practical based learning, the Cochrane 'Reviews for Africa' programme (mentioned briefly earlier in Section 6.8.1.3), was described by interviewee 8. In this initiative African health researchers and professionals wishing to conduct systematic reviews are trained in all aspects of the methodology and process. The programme supports systematic reviewers through the entire process of a systematic review. The interviewee reported that she and colleagues were in the process of evaluating the first run of the programme, but suggested the key benefits to be the 'hands-on' practical nature of the training, the chance to have dedicated time away from their day-to-day work and from interruptions, and the use of mentors.

6.8.3 Teaching quality assessment

The extent to which quality assessment was covered by training varied. In some cases the relatively short duration of training meant that quality assessment and other topics could not be explored in detail. For example, during the three day training course mentioned above by interviewee 17, the principles underpinning quality and bias were discussed, but participants did not have the opportunity to practice appraising a scientific paper. She mentioned that critical appraisal was one of the issues that participants often wanted to learn more about, and in her opinion more time would be devoted to it if were possible.

"So often we get some feedback here because they would really like to spend more time working with tools, doing critical appraisal. I mean to me that's a course in its own right, you know so. So we spend some time covering that but not the amount of time that we could"

In contrast, other interviewees reported a more thorough and 'hands-on' approach to learning about quality assessment (e.g. interviewee 13).

6.8.4 Challenges and successes

The interviewees were asked *'Are there any issues that people tend to find difficult to get to grips with?'* Table 34 shows a classification of the issues mentioned.

Table 34 - Issues that trainees find difficult to understand (sub-set of 13 interviewees who provided training)

Issues	Number (%) of interviewees
Statistical analysis	7 (54)
Quality assessment	4 (31)
Principles of evidence-based health	2 (15)
Reading and synthesising evidence	2 (15)
Study designs and hierarchy of evidence	1 (8)
Qualitative data	1 (8)

NB. Interviewees could mention more than one issue, hence total numbers exceed 13

6.8.4.1 Statistical analysis

The most common issue that trainees found difficult to understand was the statistics used in systematic reviewing, as mentioned by seven (54%) interviewees. The difficulties fell into three categories. First, two of the interviewees commented that it was difficult to describe the general principles of meta-analysis without getting into the complexities of the statistics involved. They felt that this was daunting for the trainees, particularly novices and particularly in the context of short training courses where there was little time to discuss the issues in more detail.

Second, one interviewee (10) mentioned that trainees find it difficult to recognise which particular statistical tests are appropriate for which scenarios. She described challenges in discussing the most appropriate tests to use when meta-analysing continuous (non-binary) data:

"Where there's a lot of information you can't, it's very difficult to talk about continuous data without telling people what the complexities and the dilemmas are…and that's quite complicated to describe for people who are just walking into a session"

Third, discussing how the results of statistical tests can be interpreted often caused confusion. One interviewee (4) commented on how an element of subjectivity is involved when interpreting results of a meta-analysis, and that this was a difficult issue to explain:

"So when is it causal? And when is it important? And the separating out of important from causal is tricky, and it is teaching that informed judgement, and that's a real, practice I think. So at the end of the day there's no, you know, what you can't, there's no external judge of whether something is causal or not causal, you really do need to make, to make a judgement based on a lot of things. I think that's a, that's a hard thing to teach"

6.8.4.2 Quality assessment

Another issue that presented difficulties for some trainees was quality assessment, as mentioned by four (31%) interviewees. In common with the issue of statistical analysis, described in the previous section, the challenges ranged from the conceptual to the practical.

Thinking critically

One of the problems mentioned was that trainees are not always aware of the need to think critically about research. Interviewee 6 commented on this:

"And students, especially when they first come into post-graduate training, and a lot of our students have not been studying for many many years… they're kind of blinded by the whole thing, and so they don't, they don't have an ability to understand what they're not noticing….they take everything far too much at face value…or the name on the paper is famous so it must be good"

Trainees' background

An academic grounding was mentioned as being an advantage when learning critical appraisal. For example, interviewee 11 described the Masters degree in Public Health course that she had taught. Students were given evaluation study reports to take home and to appraise. She commented that some found this easier to do than others, depending on their background. All students who do a nursing degree in her region take a critical appraisal skills module in their third year. Those that go on to do post-graduate courses such as public health have an advantage over those who have come from a different route, who tend to take longer to learn.

This comment was echoed by interviewee 9 (in response to an earlier question) who mentioned that, from her experience, one of the most common difficulties experienced by her students was understanding the differences between different evaluation designs. She commented that this was a particular problem for those who don't have a background in statistics or research.

Using judgement

There were mixed views on the utility of structured instruments in helping people to develop quality assessment skills. One interviewee (16) felt that some instruments are overly structured and lull users into the false perception that they don't have to use their judgement:

"*I mean it's the big issue often with critical appraisal is finding the things, the tools to use….And then appreciating that that's not going to replace your judgement and there's not going to be some tick box approach"*

This issue was echoed by interviewee 4 in response to an earlier question. She was concerned about an overly procedural approach to reviewing, devoid of a good understanding of the underlying methodology and any consideration of whether the results are externally valid:

"I think you've got to avoid 'the cookbook', and I think that's where Cochrane falls down and views into the system is people who are following the cookbook without understanding, you know, where to put the point estimate, down the wrong side of the line or risk and harm and benefit"

Talking about her experiences of providing training on systematic reviews to medical students the interviewee felt that this reliance on procedure was a reflection of the culture of the medical profession:

"I think the undergraduate medical students that I teach only want to know what's in the exam and they are still moving into a hospital system where it has, where what counts is the word of the consultant. And as a non-clinician methodologist they're very relatively dismissive of anything that I would say, and they get into the old cookbook medicine critiques which is frustrating"

Another interviewee (6) was more positive about the contribution of structured instruments as a way of developing critical appraisal skills. In her opinion having a framework was an effective way of helping people to think critically, particularly people who are not accustomed to doing so:

"I quite like frameworks to give students because it gives them a kind of a tool to start thinking, and many of our students haven't actually had an original thought in many years and that's

because they work in places where they not required…So giving them some tools for disentangling their thoughts is quite helpful"

Efforts to help people intimidated by the technical aspects of quality assessment to develop their skills were described. Interviewee 16 remarked that trainees attending her course are encouraged to start by applying the skills they use to weigh up the strengths and weaknesses of evidence in everyday life to research evidence:

"The course we're currently developing is on quality appraisal in fact, and this came out of an interest in how kind of existing quality appraisals are often very technical and threatening to people who were not particularly research literate. So the course we've developed is not about a kind of critical appraisal, its about how to read research, and it goes right back to first principles if you like…about the kind of quality judgements we make in everyday life, and how quality is intrinsic…So trying to introduce it as a skill we all have, and work from there towards kind of tools and techniques"

6.8.5 Availability of training and support

The interviewees were asked *'Do you think the training currently available adequately addresses the issues most relevant to health promotion?'* Only one (8%) interviewee said yes, whilst five said no (38%), and the remainder (7; 54%) were unsure. Those in the latter category felt that they did not know the area of health promotion and public health well enough to know whether there were adequate training opportunities.

Of the five who thought training was inadequate, interviewee 10 commented that this was not necessarily a problem unique to health promotion. Rather, certain health promotion interventions are regarded as being complex, in the same way that interventions on the organisation and delivery of health care, for example, are similarly complex:

"But I do think that health promotion and public health reviews need those type of reviews which tend to be complicated. I think most of the standard courses have been doing extra modules to teach people what they need to know to do complex reviews….So it's not just complex review, it's not just health promotion and public health, it's complex reviews on any topic. For example, in Cochrane the EPOC style reviews, or psycho-social reviews. You know I think there's a lot of things that come under a similar umbrella"

She commented that although extra training modules on complex reviews are available, she nonetheless considered this inadequate, and that specialist workshops are not accessible to all:

"But no, I don't think that the current training as it exists adequately meets those needs. Not everybody can get to Cochrane Colloquia, where those things are provided"

6.8.6 Suggestions for improving training and support

The interviewees were asked *'Do you have any suggestions for improving the way in which people are trained and supported to do systematic reviews /assess quality?'* Table 35 lists a classification of the suggestions made.

Table 35 - Suggestions for improving training and support / factors that facilitate effective training and support (sub-set of 13 interviewees who provided training)

Suggestions	**Number (%) of interviewees**
Format and delivery of training	
Practical exercises	3 (23)
Concentrated teaching, with breaks for practical exercises	2 (15)
More use of written materials and reading	2 (15)
Assessment and examination	2 (15)
Pitching training at the right level	1 (8)
Training content	
More coverage of complex interventions and non-experimental designs	2 (15)
More coverage of qualitative research	1 (8)
More coverage of the context of evidence based health	1 (8)
More coverage of statistics	1 (8)
Accessibility and provision of training	
Securing time and funding to undergo training	4 (31)
Accessing training	1 (8)
Training providers	
Mentorship	5 (38)
Background and skills of trainees	
Having an academic / statistical background	3 (23)

NB. Interviewees could make more than one suggestion, hence total numbers exceed 13

In many cases the suggestions made arose from the challenges reported previously (see Section 6.8.4). The suggestions were categorised into higher order themes, relating to the format and delivery of training, the content of the training, the accessibility and provision of training, the training providers and the background and skills of trainees. Each of these are described in turn in the following sub-sections.

6.8.6.1 Suggestions for format and delivery of training

Three (23%) interviewees suggested that effective learning should involve a strong element of practical experience.

Of these, interviewee 4 commented that short courses that lack any practical activities may not be adequate:

"Well I do think you have to practice it, and I do think that just saying, ah I've done a short course on systematic reviews and therefore I can do them. I think that you don't really understand bias and confounding and study design from a few short courses, unless, I really think you have to have some practice"

Interviewee 16 agreed with the need for people to receive training and support whilst doing a systematic review, but commented that some of the people who seek training do not necessarily intend to do a review:

"Well the consensus seems to be the way to train to do people to do reviews is the more kind of, is throughout the process isn't it?... That seems to intuitively make sense to me. While the sort of training we do is a whole step back from that, before people are, you very rarely get people coming along courses who are starting a review, they're thinking about, as maybe something to add to their methodological toolkit"

6.8.6.2 Suggestions for accessibility and provision of training

Four (31%) interviewees commented that there would be greater uptake of training if more time and funding was available. It was suggested that this would enable people to take periods of time away from work necessary to learn systematic reviewing. The funds would cover the costs of the training course itself, plus costs to cover their absence from work. Interviewee 14 commented that encouraging health professionals to learn how to conduct systematic reviews is desirable, but constrained by a lack of time and funding:

"I mean the reality is that if you're a full-time health professional you haven't got time for a systematic review, much as we'd like it to be done by full-time practising health professionals because we might get different questions and different answers. And maybe that's one of the things that we need, is more investment in buying out health professionals to do things like this, to give them the time and the support to do it"

6.8.6.3 Suggestions for training providers

Five (38%) interviewees suggested that learning to do systematic reviews could be improved with the use of mentors. This was influenced by the fact that, as mentioned earlier, many of the interviewees had found mentors to be very helpful in their own learning (see Section 6.7.2). One of the interviewees (12) viewed mentorship as being complimentary to other forms of learning:

"You get to a stage where you can read the textbooks, you know or you can be directed to a manual or a handbook, but then what you really need is, I think is some guidance or mentorship over that whole process just to help you think about what you're doing and to help you apply, you know what you've read in the book to the real scenario"

6.9 Chapter Summary

This chapter has presented the findings of the second stage of the research, the semi-structured interviews with a sample of systematic reviewers in health promotion. A lot of ground has been covered, including: the interviewees' views on the strengths and weaknesses of systematic reviews, the challenges they have faced and the strategies they have used to deal with them, their experiences of conducting quality assessment and views on quality assessment criteria, and their experiences of learning and teaching skills for systematic reviewing. The next chapter discusses these findings in detail, together with the results of the first stage of the research.

Chapter 7 - Discussion of findings

Chapter outline

The purpose of this chapter is to discuss the implications of the findings of this thesis. It draws together the key results of Stage 1 of the research (Chapter 4) and Stage 2 (Chapter 6). Each of the research objectives proposed in Chapter 1 are revisited in the light of the findings and recommendations are made for policy, practice and research throughout the chapter in **bold type**.

7.1 Strengths and weaknesses of systematic reviews

Research objective:
To assess views on the strengths and weaknesses of systematic reviews of health promotion

The interviewees in this study believed that systematic reviews have many of strengths, but they also acknowledged that they suffer from a number of weaknesses (Chapter 6). The impression given by their comments was that they considered the weaknesses to be out-weighed by the strengths, and overall they retained their support for systematic reviews.

7.1.1 Rigorous and transparent methodology

The most commonly mentioned strength of systematic reviews was their rigorous and transparent methodology, cited by just under half of those interviewed. However, the interviewees acknowledged that the methodology must be applied correctly, to avoid biases arising. It was also appreciated that despite efforts to be systematic, inevitably biases do occur and sometimes reviewers need to exercise their own judgement. That systematic reviews are wholly transparent is not a view accepted by all. For example, Hammersley (2006) draws on the philosophy of Michael Polanyi (1966) to suggest that 'tacit knowledge' (knowledge that people carry in their minds that cannot easily be communicated), plays a key role in all science. That systematic reviews are exempt from this assumption is not tenable, in his opinion.

Ogilvie *et al* (2005: 891) suggest that judgement is often required and complete transparency is unrealistic:

"The evidence never speaks for itself, but is always open to interpretation, and there are elements of the review process that entail judgement and cannot be made entirely transparent or replicable"

The need to exercise judgement was a common theme arising from the interviews and is discussed throughout this chapter.

7.1.2 Informing the research agenda

Another strength mentioned in the interviews was that, where gaps in the evidence base exist, systematic reviews can be a tool for identifying the gaps and advocating further research. Systematic reviewers are therefore in a good position to inform the research agenda given their detailed assimilation of the evidence base. However, at least one interviewee questioned whether there was a process for disseminating research recommendations from reviews, and whether there is much incentive for funding high quality health promotion research at all. Although the Wanless report into public health and health inequalities noted an historical lack of funding for public health intervention research (Wanless, 2004), these misgivings may no longer be necessarily founded as increased funding for research was one of the policy imperatives in the Choosing Health white paper for public health in England (Department of Health, 2004).

There is some evidence to show that the commitments made by Wanless and the Department of Health have been fulfilled. Mechanisms are in place in the UK for routinely extracting research recommendations from systematic reviews so that they can be considered as potential research priorities. For example, the Department of Health supported NIHR Health Technology Assessment Programme routinely scans the research recommendations made by Cochrane reviews, as well as reviews it commissions itself, as part of its identification and prioritisation process (Allen and Stockley, 2008). Topic suggestions relevant to health promotion and public health are considered by its Disease Prevention Panel (DPP). Since the inception of the DPP in 2004, a number of health promotion topics have been commissioned by the programme, demonstrating an increase in research funding in this area. Furthermore, Public Health Guidance issued by NICE, which is based on systematic review, routinely make recommendations for research (National Institute for Health and Clinical Excellence, 2006).

It is evident that an infrastructure does exist for identifying gaps in evidence base that have been identified by systematic reviews, and translating those gaps into commissioned research projects. It is unlikely, however, that all systematic reviews will have the potential to influence

the research agenda in this way, particularly those in the 'grey' (unpublished) literature, or published in journals not indexed by electronic bibliographic databases which are harder to access (e.g. Medline; Embase).

7.1.3 Supporting policy and practice

The interviewees also considered that one of the strengths of systematic reviews is the ability to enable policy makers and practitioners to make better decisions. In particular, it was suggested that reviews have the greatest potential for change when conducted as part of a policy making initiative, an example being the US Center for Disease Control and Prevention Taskforce on Community Preventive Services (Anderson *et al*, 2003). It cannot be assumed, however, that there is a direct link between the findings of systematic reviews and policy. As one interviewee commented, policies are often implemented with scant regard for the evidence, an assertion backed up by examples from the literature (Petrosino *et al*, 2002).

What evidence, therefore, is there that systematic reviews of the effectiveness of health promotion have a positive impact on decision making? There does not appear to be an extensive empirical evidence base on the impact of systematic reviews. However, there are examples of policy-orientated systematic review programmes which demonstrate a transparent link between evidence and policy. For example, in England and Wales systematic reviews and economic evaluations conducted for NICE's Centre for Health Technology Evaluation inform decisions about whether or not health technologies (mostly pharmacological treatments) are recommended for use in the health service. NHS trusts are mandated to implement the guidance within three months (Department of Health, 2003; National Institute for Clinical Excellence, 2004) (Although there is evidence that uptake of guidance by the NHS has been variable (Sheldon *et al*, 2004)). Moreover, an evaluation of the first ten years of the NIHR HTA Programme concluded that it had considerable impact in knowledge generation, policy and to some extent on practice (Hanney *et al,* 2007). However, this evaluation pre-dated the expansion of the programme's remit to cover health promotion with the establishment of the DPP in 2004. Therefore, a recommendation from this research is for further evaluation of the impact on decision making of systematic reviews of the effectiveness (e.g. the NIHR HTA DPP).

7.1.3.1 Systematic reviews of relevant topics?

The contribution of reviews to informing policy and practice depends, in part, upon whether the scope of the review is meaningful and relevant to its intended audience. One of the interviewees' key concerns was whether or not systematic reviews always achieve this. They suggested that reviews are often conducted where it is perceived that enough evidence exists to

make a review worthwhile, or where it is anticipated that there will be better quality evidence. These topics may not necessarily be the most relevant to those in the field, which has been a criticism of systematic reviews for some time (Whitehead, 1996; Fraser, 1996). Furthermore, where reviews have not been conducted the gaps that exist in the evidence base may go unnoticed by research commissioners, negating what the interviewees suggested to be one of the key strengths of systematic reviews. This has been referred to as the 'inverse evidence law', whereby the least is known about the interventions likely to be of most importance in terms of promoting health (Nutbeam, 2003; Ogilvie *et al*, 2005; Petticrew *et al*, 2004). Policy makers have argued that much public health research, particularly in the area of health inequalities, is of little relevance to them. Although not specifically a criticism of systematic reviews, they note the prevalence of research that does not answer policy questions, what they call 'policy-free evidence' (Petticrew *et al*, 2004).

It would be misleading, however, to conclude that systematic reviews necessarily fail to address policy-relevant topics. There are published examples of reviews that seek stakeholder input throughout the process to ensure the review answers the most relevant questions. For example, the review of sexual health promotion interventions for men who have sex with men that I was involved in (as mentioned in the Introduction to this thesis) was advised by a panel of practitioners, researchers and policy makers who helped define the scope (Rees *et al*, 2004a; Rees *et al,* 2004b). (The use of advisory groups is discussed in greater detail later in this chapter, see Section 7.9.3)

7.1.4 Summary and recommendations

This research has identified views on some of the strengths and weaknesses of systematic reviews. The perceived benefits include rigorous methodology, the ability to identify where primary research is needed, and the contribution to effective decision making. However, doubt was cast over the ability of reviews to always address policy-relevant topics. **Whilst there is some evidence to support the assertions made, it is recommended that the impact of systematic reviews of the effectiveness of health promotion be evaluated on an on-going basis to ensure that the needs of all stakeholders are being met. This is particularly important, in terms of public accountability, given government commitments to increase funding for evidence synthesis in this area.**

7.2 Challenges facing systematic reviewers

Research objective:

To assess the challenges people have faced when doing systematic reviews of health promotion?

- How have these challenges been dealt with?
- With what success?

The interviewees mentioned a number of challenges they had encountered whilst conducting systematic reviews. The key challenges were a lack of time and resources to do reviews, and the complexity of the evidence.

7.2.1 Time and resources

As reported in Chapter 6, the most common challenge, mentioned by just over two-thirds of those interviewed, was a lack of time and resources for systematic reviewing. This was a recurring theme in the interviews, and is discussed throughout this chapter. Systematic reviews of public health and health promotion topics were considered to be more time consuming than those of clinical topics, on account of the often large volumes of evidence to process.

7.2.1.1 Securing funding

The time required to put together a successful grant proposal for a review was also a barrier, even for those with an academic position. It was also perceived that adequate funding opportunities for systematic reviews were scarce. The solution to the problem of securing funding, for some, was to self-finance systematic reviews, doing them largely in their spare time on a 'shoe-string' budget. Whilst this may have been feasible in the past the feeling now was that, given the increased expectations placed on the conduct and reporting of systematic reviews (Centre for Reviews and Dissemination, 2009; Higgins and Green, 2008; Moher *et al*, 2000), and the current academic climate with its focus on publishing research articles in high ranking peer-reviewed journals (Godlee, 2006), this would no longer be realistic. Oakley (2003) suggests that the average cost of a producing a systematic review is between £75,000 to £100,000. Given inflation since this estimate was made the average cost is now likely to be between £100,000 to £150,000.

Interestingly, Nind (2006) in her reflexive account of systematic reviewing in education, noted that one of the incentives for doing a systematic review was because there was funding

available. She later notes, however, that the funding given did not match the effort required to conduct the review, leading to pressure on the review team. It would seem, therefore, that the challenge for systematic reviewers is convincing research commissioners that *adequate* funding is provided.

There are some encouraging signs that provision of funding is on the agenda, given the aforementioned expansion of funding for health promotion evidence synthesis by the NIHR Health Technology Assessment Programme (Allen and Stockley, 2008). Furthermore, the NIHR Public Health Programme (PHR), launched in August 2008, will spend £10 million a year on primary and secondary research into the effectiveness and broader impact of multi-disciplinary interventions to promote health and reduce financial inequalities (National Institute for Health Research, 2008; Walley and Thakker, 2008). A commitment to systematic reviews was also made in the UK Government's national health research strategy 'Best Research for Best Health' (Department of Health, 2006). Whilst these are encouraging signs of investment into high quality evidence synthesis it will be important to monitor how the money is spent, in terms of ensuring reviews are realistically funded, and whether there is an equitable distribution of funding opportunities across all aspects of health promotion and, more broadly, public health.

7.2.1.2 The opportunity cost of doing systematic reviews

Traditionally, the Research Assessment Exercise (RAE) (which evaluates research done by UK higher education establishments in order to determine future research funding) has favoured 'pure' science over applied practice-orientated research (Godlee, 2006) (although for the first time systematic reviews were recognised by the 2008 RAE, under the Health Services Research Panel). Academics will no doubt have felt pressured to pursue other activities that might yield higher credit in a shorter time scale. There may be an opportunity cost associated with conducting a time-consuming and labour intensive systematic review, particularly with limited funding. Some of the most prestigious medical journals, in terms of their impact factor, do not routinely publish systematic reviews (Brown, 2007). For example, the New England Journal of Medicine had an impact factor of 51.29 in 2006, compared with the British Medical Journal (which does publish systematic reviews) whose 2006 impact factor was 9.24. Instead, academics may be encouraged, or in some cases, obliged to spend their study time publishing papers in high quality journals based on primary empirical or theoretical research. Greater acceptance of systematic reviews by high quality academic journals in health, in terms of editorial policy, would facilitate effective dissemination of their findings, and encourage more researchers to consider producing them.

The increasing complexity of the methods for systematic reviewing, coupled with higher expectations of their quality, means that reviews conducted using limited resources may fall short of the mark by current standards. Only reviews that are adequately funded and resourced are likely to be considered credible. The upshot of all this is that there may be significant disincentives for people to consider conducting systematic reviews, particularly those with limited time and resources, but who by virtue of their expertise in a particular topic, may be in a particularly appropriate position to do so (see Section 7.9).

7.2.2 Complexity of the evidence

The complexity of some of health promotion interventions, and the observational evaluation designs sometimes used, was another key challenge the systematic reviewers interviewed faced. They commented on the difficulties associated with analysis of cluster trials designs, and mentioned a perceived lack of support and guidance to review complex interventions. It was also noted that some existing frameworks for systematic reviews, such as that used by Cochrane, were inappropriate for topics other than straightforward clinical interventions.

To overcome these difficulties the interviewees sought advice from others, or consulted methodological guidelines, with varied success. There has been increasing interest in the evaluation of complex interventions over recent years, and what appears to be an evolving methodology. For example, in 2000 the Medical Research Council published a framework on the evaluation of complex health interventions (Campbell *et al,* 2000) (updated in 2008 (Craig *et al*, 2008)). They proposed a phased approach to evaluation of such interventions, and this has subsequently been applied by others (Campbell *et al,* 2007; Murchie *et al,* 2007). Furthermore, Pawson (2006b) proposed five strategies for evaluating complex interventions (such as Health Action Zones and New Deal for Communities) There are also recent examples of process evaluation conducted within RCTs to shed light on the contributory factors in the effectiveness (Oakley *et al,* 2006; Stephenson *et al*, 2004) or failure (Elford *et al*, 2002) of complex health promotion interventions. These examples suggest increased intellectual investment into methods of assessing and analysing complex interventions, and better reporting in primary studies. This will hopefully lessen some of the problems faced by systematic reviewers in this area, but no doubt further methodological research needs to be done.

7.2.3 Summary and recommendations

The results of this study show that systematic reviews, particularly of complex health promotion interventions, can be time consuming and costly to conduct. Progress seems to have been made in terms of investment in an adequate infrastructure and capacity to produce reviews. **It is**

recommended that investment be maintained, to ensure that systematic reviews are commissioned and produced in a timely fashion and to a high standard. Funding opportunities for systematic reviews should be evaluated to ensure they are accessible to all, and that they represent the diversity of health promotion interventions and topics.

There may be disincentives, particularly for academics, to consider producing systematic reviews. **It is recommended that academic journals, including those with a high impact factors, consider accepting systematic reviews for publication as part of their editorial policy.**

7.3 The extent to which quality is assessed

Research objective:

To assess the extent to which systematic reviews of health promotion assess the quality of included studies:

- What are the barriers to, and facilitators of, quality assessment?

The results of this study show that quality assessment is an integral feature of systematic reviews. Quality assessment was classified as being performed in 93% of the reviews assessed in part one of the study (Chapter 4), and all of those interviewed in part two reported that they routinely assess quality (Chapter 6). The views of the interviewees are best encapsulated by the quote from one of them that:

'The majority of the mainstream systematic reviewers agree that quality assessment is one of the things that underlines it' (interviewee 10)

These results resonate with those of a larger survey of systematic reviews of a range of different types of interventions (including health promotion), published between1995 and 2002 (Moja *et al*, 2005). The study found that quality assessment was performed in 88.5% of systematic reviews. However, these findings are in contrast to those of an earlier study by Peersman *et al,* (1999) (Chapter 1), where only a third of the 400 health promotion reviews sampled assessed quality. Although the present study used a much smaller sample of reviews, it is more up to date than Peersman's and is augmented by the comments made by the interviewees. Together with the results of Moja *et al* (2005) the results suggest a more optimistic picture, with quality assessment now commonplace in systematic reviews of health promotion.

7.3.1 Poor reporting of primary studies

Despite the widespread assessment of quality, the interviewees mentioned barriers which inhibited this process (Chapter 6). Again, time was an issue. It was commented that quality assessments were not always as detailed as would have been desired due to pressures to complete the review on time. This relates to one of Hammersley's (2006) criticisms of systematic reviews, namely, that the quest for exhaustiveness, particularly in literature searching, is at the expense of time needed to read, think and reflect on the literature. However, an alternative view might be that although systematic reviews are often time consuming to conduct, exhaustiveness and attention to detail are not mutually exclusive. Rather, adequate planning and management should ensure that tasks which are usually performed later in the process (e.g. quality assessment) are not jeopardized, in terms of time, by tasks done earlier on (e.g. literature searching).

Linked to problem of limited time were the challenges presented by poorly reported studies, which the interviewees mentioned were more laborious to appraise. To overcome the problem some reviewers contacted the authors of the studies to clarify ambiguities and request missing data, with varying degrees of success. One interviewee described an innovative approach, whereby their draft assessment of quality was shared with the author of the study who commented on accuracy and fairness. One of the advantages of this approach is that it may facilitate a more accurate assessment of quality.

However, it could be argued that the reviewers may be influenced, implicitly or explicitly, by the author to provide a favourable judgement. As systematic reviews should be independent of all competing interests (Cochrane Collaboration, 2008) it could be suggested that reviewers and authors have minimum contact, particularly if the authors have commercial or other affiliations with the producer of the intervention (Lexchin *et al*, 2003). On a practical level it may also introduce a further step into the systematic review process, requiring more time and resources and driving up the cost of reviews even further. As one interviewee mentioned, there is not always time to contact authors even for basic clarification.

The implications of this is that, where time is pressured, poorly written studies are likely to be inadequately, and potentially inappropriately, judged in terms of their quality. Reviewers may be forced to conclude that quality is 'unclear' whereas, in reality, it might be sound or unsound. A recommendation from the current study would be to allow extra time for 'decoding' poorly written studies and, where necessary, for contacting authors. However, it is acknowledged that this may not always be realistic, and systematic reviewers would have to convince funders that

the extra time needed would be cost-effective, in terms of a more accurate and just review. Further empirical investigation into methods for successfully dealing with poor reporting would therefore be valuable.

The problem of poor reporting will hopefully lessen through the impact of initiatives to improve the standard of reporting in health research. In particular the CONSORT statement (Consolidated Standards of Reporting Trials) (Altman *et al*, 2001; Moher *et al*, 2001a), which aims to improve the standards of reporting in RCTs, has recently been extended to incorporate RCTs of non-pharmacologic interventions, including behavioural interventions (Boutron *et al*, 2008). However, empirical evaluations of the impact of the CONSORT statement on the quality of reporting in RCTs have yielded mixed findings. Some studies, including those conducted by the team who devised the statement, have reported an increase in reporting standards (Moher *et al,* 2001b; Plint *et al,* 2006). Similarly, Devereaux *et al* (2002) and Mills *et al* (2005) both reported an increase in standards but noted that compliance with some of the CONSORT recommendations was not universal. Evaluations of the statement in specialties such as palliative care (Piggott *et al*, 2004) and alternative therapies (Liu *et al*, 2003) found that reporting remained poor. This is in accordance with the findings of Altman *et al* (2005) who found that uptake of the statement by academic journals was higher in the area of general medicine (53%) than in specialties (18%). He also found that instructions to authors were often ambiguous and referred to out of date versions of the statement.

It is important that journals specializing in health promotion and public health sign up to the CONSORT statement, and other relevant initiatives such as the TREND statement (Transparent Reporting of Evaluations with Nonrandomized Designs) which was devised specifically to raise standards of the reporting of non-randomised studies in public health (Des Jarlais *et al*, 2004). At present it appears that only relatively small number (<20) of academic journals have signed up to the latter, and there do not appear to be any published evaluations. The impact of these statements on the reporting of health promotion evaluations should be rigorously assessed.

7.3.2 Using a good quality assessment instrument

Use of a good instrument was considered to facilitate the process of quality assessment, as mentioned by just over half of those interviewed (Chapter 6). A 'good' instrument was considered to be one that covers all of the attributes of quality relevant to the topic being reviewed, and one that is clearly planned and justified in advance. Although a valid and reliable quality assessment instrument was considered to be helpful, particularly when learning how to appraise studies (as discussed later in Section 7.11.2), it was noted by a couple of interviewees

that some form of personal judgement is nearly always required. They suggested that this may potentially introduce bias into the process, but this was considered inevitable. As mentioned earlier in Section 7.1.1, the issue of reviewer judgement and subjectivity was a recurring theme in the interviews in this study.

This appears contradictory to the key tenets of evidence synthesis: that the process should be as objective, transparent and reproducible as possible (Centre for Reviews and Dissemination, 2009; Egger *et al*, 2001). Hammersley (2001) is critical of the fact that systematic reviews unquestioningly adopt a positivist model of science and a commitment to procedural objectivity. He argues that much research cannot be reduced to following a set of rules, and that the principles of systematic reviewing largely ignore the criticisms made of positivist science over many decades. However, others would argue that this does not preclude that systematic reviewers, whilst committed to minimizing bias, can be mindful of their subjectivity and take steps to try and account for it (Pawson *et al*, 2005; Petticrew *et al*, 2004; Nind, 2006). The suggestion that a blithe acceptance of positivism underpins the theory and practice of systematic reviewing may not be tenable.

Hammersley (2001) also takes issue with the practice of quality assessment in general, particularly the assumption that it can be done by routinely applying a set of criteria, based on a standard hierarchy of evidence, without consideration of the context in which a study was done. This view is even shared by some who have conducted systematic reviews. For example, the EPPI-Centre facilitates the production of systematic reviews of the effectiveness of education by external review groups (mostly comprising professionals working in the education sector). Oakley (2003), in a reflective account of the reviews conducted to date, noted that one of these review groups did not offer judgement on study quality as they disagreed with the positivist assumptions underpinning the theory and practice of quality assessment. This issue is expanded upon later in this chapter (Section 7.9.2).

7.3.3 Summary and recommendations

This study has found that, in common with other recent investigations, quality assessment is commonplace in systematic reviews of health promotion. It has also illuminated some of the barriers and facilitators to the process of quality assessment. Notably, time pressures can impede a thorough assessment of quality, a problem that is exacerbated by poorly reported studies. **It is therefore recommended that systematic reviews are realistically planned so that adequate time is available to ensure a thorough and fair assessment of the quality of the evidence.**

This study has also elicited examples of innovative strategies to obtain accurate information from study authors upon which to base a quality appraisal. **Such strategies should be subjected to empirical evaluation to assess the benefits in terms of a potentially more thorough and rigorous systematic review, and to establish whether strategies are realistic in terms of the time and resources required.**

Initiatives to improve standards of reporting of primary evaluations, such as the CONSORT and TREND statements are in place. **It is recommended that academic journals specializing in health promotion and public health sign up to these statements. The effectiveness of these initiatives should be monitored on an on-going basis.**

7.4 Quality assessment criteria

Research objective:

To assess the criteria that systematic reviews of health promotion use to assess the quality of included evidence:

- Which criteria are used?
- Why have these criteria been chosen?
- Do these criteria address acknowledged threats to internal validity?

When asked which criteria they use to appraise studies, most of those interviewed mentioned a selection of the key dimensions of methodological quality, as opposed to a complete list of all criteria (Chapter 6). These included attrition and loss to follow-up, the validity and reliability of data collection and analysis methods, blinding, and methods of allocation to study groups including concealment of allocation. The dimensions reported were generally in accordance with the criteria employed by the systematic reviews mapped in this study (Chapter 4). The remainder of this sub-section focuses on the criteria used in those systematic reviews.

7.4.1 Criteria relevant to controlled trials

Randomised and quasi- or non-randomised controlled trials were permitted in the vast majority (96%) of systematic reviews mapped in this study (Chapter 4). About half of the reviews restricted inclusion to RCTs and experimental designs. This finding is at odds with the long-standing argument that RCTs are inappropriate, or not feasible in health promotion (Nutbeam, 1999; Speller *et al*, 1997; Tones *et al*, 2000; Tones and Tilford 2001). It shows that RCTs of health promotion are possible (although whether they have been conducted to an acceptable

standard is a separate issue discussed below). However, health promotion systematic reviews do not exclusively favour RCTs, as at least half permitted inclusion of observational designs.

The mapping found that the most commonly assessed quality criteria relevant to controlled evaluation designs included the adequacy of the method of randomly allocating people to intervention or comparison study groups (57%), and the comparability of the study groups at baseline (50%) (both of which protect against selection bias), as well as blinding (which protects against performance bias) (43%).

In terms of blinding, the reviews tended to assess whether or not the outcome assessors, rather than the intervention recipients, were unaware of their intervention allocation (protecting against detection bias). A potential explanation for this is that it may be more feasible to mask treatment assignment from outcome assessors rather than the intervention recipients (Stephenson and Imrie, 1998). For some types of outcome measure the assessors may have no or minimal contact with the recipients (e.g. laboratory pathologists analyzing blood samples to assess biochemical changes following an intervention) (Boutron *et al*, 2007; Flay, 1986).

Blinding was not discussed in detail by the interviewees in this study, but at least one person associated it with 'standard' and 'rigid' models of quality assessment, such as that used by the Cochrane Collaboration, and something that was not possible in health promotion (Chapter 6). While this may have some grounding, it may not be a universal view. At least one of the reviews mapped in this study reported that, despite performance bias being an inherent problem in health promotion, blinding of participants was possible for some interventions (Dyson *et al*, 2005). Dyson *et al* (2005) remarked that, in one of their included studies, it was possible for mothers receiving special packs to promote the uptake of breastfeeding to be blinded to whether or not they received the experimental or the control (placebo) pack. It is evident, therefore, that blinding is possible in some health promotion interventions, and is a criterion that some systematic reviewers appraise studies by.

7.4.2 Criteria relevant to evaluation designs in general

The mapping found that the most commonly assessed quality criteria applicable to evaluation designs in general (i.e. not just experimental designs) were attrition (86%), the validity and reliability of data collection instruments/methods (64%), and the validity and reliability of data analysis (57%). As discussed in Chapter 1, attrition is a particular issue in the evaluation of health promotion, particularly where outcome follow-up is long term, and also where study populations are transient (Coyle *et al*, 2006; Gwadz and Rotheram-Borus, 1992). This is

reflected by the results of this study which show that the vast majority of reviews included attrition as a quality criterion. Some of the reviews also assessed whether studies adequately attempted to compensate for attrition (e.g. 'intention to treat / intervene' analysis), as classified under the category of the validity and reliability of data analysis.

In terms of the validity and reliability of data collection instruments/methods, nearly two-thirds of the reviews assessed whether steps were taken to control for bias associated with self-reported outcomes (e.g. health related behaviour), such as triangulation with physiological data. This is an issue particularly relevant to, although not exclusive to, health promotion where self-reported outcomes are commonly measured in lieu of physiological outcomes which may not manifest themselves for many years (Tones and Tilford, 2001).

7.4.3 Justification for the criteria used

Just under three-quarters of the reviews mapped in this study reported some justification for the quality criteria they employed. However, the justifications were not always explicit. The reviews tended to cite other reviews in which the criteria (or iterations of the criteria) had been employed, without reporting why they had chosen, or any discussion of their strengths and weaknesses. On occasion the justification was a little more explicit, with reference made to the fact that the criteria adopted were 'high profile' (Huibers *et al*, 2003). The implication of this is that if the criteria are widely used they are therefore credible. However, whether this assertion is tenable is open for question.

Only two-fifths of the reviews mapped justified their choice of criteria by citing empirical texts on risk of bias. A similar finding was reported by Deeks *et al* (2003) in their review of quality assessment instruments for non-randomised studies. They reported that just 73 (37%) of the 193 instruments surveyed were developed according to the methodological literature, or consensus. In the majority of cases the rationale for the criteria was not given. However, Moher *et al* (1998) reported contrasting findings to those of the current study, and to Deeks *et al* (2003). In Moher's review of instruments used to assess the quality of RCTs of health care, it was found that the majority of the instruments included questions on the key threats to internal validity, based on empirical evidence (at the time) on the characteristics of RCTs related to bias. However, a possible explanation for the divergent findings is that Moher *et al* restricted their study to instruments used to assess the quality of RCTs (rather than all designs). Deeks *et al* (2003) suggests that there is more empirical evidence on the threats to the internal validity of RCTs and controlled trials than there is for observational studies.

In the two-fifths of reviews mapped in the current study which justified their criteria according to empirical texts on bias, none of the texts were cited by more than one review. This finding suggests variability in terms of the sources drawn on by systematic reviewers, although the reason for this is not clear. It is interesting that concealment of the allocation of participants to study groups in controlled trials was only assessed by a third of the reviews mapped (all of them Cochrane reviews), despite widely publicised empirical evidence showing that it can exaggerate study effects if compromised (Egger *et al*, 2001; Moher *et al*, 1998; Schulz *et al*, 1995) (although see below). Thus, non-Cochrane reviews do not appear to be mindful of the empirical evidence, at least as regards allocation concealment. However, the fact that attrition bias and performance bias were commonly assessed by the reviews does not suggest that empirically recognised biases were completely ignored. The issue is that the reviews did not always explicitly report why they had chosen to use these criteria. If reviews do not provide a sound and transparent rationale for their choice of criteria they may be seen by some as lacking in credibility. Similarly, whilst the interviewees in this study gave the impression that they were aware of the key threats to validity in evaluation, only two mentioned that their choice of criteria had been informed by empirical evidence of bias. It is therefore recommended that systematic reviews explicitly report the basis for their choice of criteria, preferably with reference to high quality empirical evidence on risk of bias.

Interestingly, one interviewee perceived that the findings of some empirical studies are conflicting and can be open to differences in interpretation. This perception has some grounding. Reviews of systematic reviews (i.e. tertiary reviews) that sought to compare the effects of randomised and non-randomised controlled trials have concluded that the results between these designs can sometimes be divergent. In some cases effects were similar between randomised and non-randomised designs, and in other cases they were different, with no consistent pattern in effect sizes (Deeks *et al*, 2003; Oliver *et al*, 2008). Furthermore, returning to the issue of allocation concealment (as discussed above) recent evidence suggests that the influence of poor concealment may not be as great as previous empirical studies have shown (Pildal *et al*, 2007). The influence also varies according to whether objective or subjective outcome measures are used (Wood *et al*, 2008), complicating matters further. This raises the question that if there are uncertainties in the conclusions and interpretation of cutting edge empirical methodological research, then can there really be consensus amongst systematic reviewers on which criteria to appraise studies?

Contradictions in the findings of methodological studies also suggest the need for further empirical research to avoid some of the methodological weaknesses (referred to as 'meta-confounding') observed in the studies reviewed by Deeks *et al* (2003) and Oliver *et al* (2008).

That is, studies should use larger data sets, appropriate statistical techniques, and eliminate any differences other than the presence or absence of randomization that might confound results (where the objective is to assess the effect of randomization, that is). There is also a need for an infrastructure to enable effective dissemination of the findings to systematic reviewers, particularly to clarify issues where there appears to be potential for misinterpretation or mixed results. The Cochrane Methodology Register, one of the constituent databases of the Cochrane Library, is a repository of such studies. However, systematic synthesis of this evidence similar to that by Deeks *et al* (2003) and Oliver *et al* (2008) would be useful. Research funders such as the Medical Research Council (MRC), which recently took over the commissioning of methodological research from the NIHR, should consider supporting this work (Medical Research Council, 2008).

7.4.4 Summary and recommendations

This study has mapped the quality assessment criteria used by systematic reviews of health promotion. In terms of design the majority of reviews permitted inclusion of RCTs and controlled designs, and half also included observational designs, suggesting a pragmatic approach to quality assessment encompassing the heterogeneity of designs used to evaluate health promotion. The reviews commonly assessed the presence of key threats to bias, such as selection bias, attrition bias and performance bias. Some of the potential biases which are particularly acute in health promotion, such as the use of self-reported outcomes, were assessed, although to a lesser extent.

The interviews identified the perception that empirical methodological studies have yielded conflicting findings in terms of the degree to which certain methodological characteristics of evaluations bias study effects. **It is recommended that further empirical methodological research be commissioned, including tertiary reviews, to try and reconcile these divergent findings. These findings should underpin updated guidance to systematic reviewers (e.g. the Cochrane Handbook) to reduce uncertainty and to facilitate consensus in terms of quality assessment criteria.**

Whilst the majority of reviews justified their choice of quality assessment criteria, only a small proportion cited empirical evidence of how the criteria take into account bias. **It is recommended that systematic reviews explicitly justify their choice of criteria with reference to up to date empirical evidence demonstrating how the criteria account for bias.**

7.5 Consensus on quality assessment criteria?

Research objective:
To assess whether there is consensus on the criteria by which health promotion evaluations should be assessed in systematic reviews

To compliment the data on the use of quality criteria derived from mapping the reviews presented in the previous section, evaluative comments on evaluation designs from the published systematic reviews (Chapter 4), and from those interviewed (Chapter 6) were analysed. The key finding is that although RCTs and experimental designs are widely supported, debates still remain about their feasibility and appropriateness in health promotion.

7.5.1 Comments in support of RCTs and experimental designs

The interviewees did not generally comment in detail on the strengths and weaknesses of RCTs, but did suggest that systematically reviewing RCTs was easier than non-randomised studies. There was strong support for RCTs by the published systematic reviews mapped in this study (Chapter 4), where just under two-thirds commented on the strengths of this design. On occasion support was explicit. For example, when reviews made recommendations for further research, around half favoured RCTs and experimental designs in general. Some even provided practical advice on how to conduct them in locations where experimental evaluation is known to be problematic. However, to a lesser extent it was suggested that RCTs can sometimes be unethical, expensive and restrictive of participant choice.

The 'defeatist' attitudes of some evaluators have long been rejected by those who argue that RCTs of health promotion are feasible (Loevinsohn, 1990). As mentioned in Chapter 1, Ann Oakley points out that there is a long-standing but overlooked history of experimental evaluation particularly in the US (Oakley, 1998). She argues that the existence of this golden age of experimental evaluation is testament to the feasibility of such methods. However, what might be achievable in the US may not necessarily be possible in other locations, particularly developing countries where there may be significant social, economic and political barriers to evaluation (Pettifor *et al*, 2007) (although there are published examples of RCTs in developing countries, such as the cluster behavioural HIV prevention trial in Uganda reported by Kamali *et al*, 2003). Whilst it may be true that some evaluators are unnecessarily pessimistic about the practicability of experimental designs, it is also likely that for some initiatives, such as community development, and area wide interventions, their use will remain difficult, if not impossible (Bonell and Imrie, 2001; Hallett *et al*, 2007; Tones, 2000).

7.5.2 Critical comments of RCTs and experimental designs

This study also found that just under a quarter of the systematic reviews commented on the weaknesses of RCTs and experimental designs. They noted that, although conceptually a strong design, RCTs can be poorly conducted. For example, although randomisation is thought to protect against selection bias, by chance it may not always ensure an equal distribution of participants across study groups (Roberts and Torgerson, 1999). Flaws in the randomisation process, can occur, may also lead to baseline imbalances between groups (Kjaergard *et al*, 1999; Schulz *et al*, 1994). If adequate steps are not taken to compensate for imbalances results may be confounded by 'chance bias' (Roberts and Torgerson, 1999; Senn, 1994). Attrition, if not successfully accounted for in the data analysis, was also mentioned as something that could potentially weaken the utility of the randomised design. The results of this study show that even though RCTs and experimental evaluation designs are considered by some to be the gold standard in the evaluation of health promotion (Flay 1986; Oakley *et al*, 1995; Rychetnik *et al*, 2002; Stout and Rivara, 1989), their potential shortcomings are acknowledged by systematic reviewers. The fact that the reviews formally assessed the extent to which selection bias, performance bias, and other biases were present suggests that RCTs are not necessarily accepted on face value by virtue of their design. The general support for RCTs is therefore balanced by acknowledgement that they can be poorly executed, and systematic reviewers are realistic about their limitations.

7.5.3 Comments in support of observational studies

Despite general acceptance by the systematic reviews that observational studies have shortcomings, some of them commented on specific circumstances where they have been considered useful (e.g. in the early days of the HIV/AIDS epidemic where rapid evaluation of preventive interventions was an imperative). Their comments underscore the utility of observational evidence in circumstances where experimental evidence is lacking. Other commentators have discussed the acceptability of non-experimental evidence (Hallett *et al*, 2007; Petticrew and Roberts 2003; Thomson *et al*, 2004), and frameworks for its use in decision making have been proposed. For example, Slavin in the 1980s introduced the concept of 'Best Evidence Synthesis' (Slavin, 1986; 1995), and applied it in the field of education (Slavin 1990). In this approach, studies high in internal and external validity would be prioritised for inclusion in a systematic review. In areas where no such studies exist then evidence from studies with lower internal and external validity may be permitted as 'fit for purpose', with appropriate caveats to alert the reader to potential biases (note the similarities to the principles of the 'hierarchy of evidence' as described in Chapter 1).

Systematic reviews which adopt this perspective do not necessarily abandon considerations of quality, but seek to use the best *available* evidence (Ogilvie *et al*, 2005; Petticrew, 2003). The advantage is that these reviews will be able to make conclusions, albeit cautious ones, about the effectiveness of interventions. This is in contrast to reviews that are inconclusive due to lack of sound evidence, what Petticrew (2003) refers to as the 'stainless steel law' of systematic reviews, namely, that the more rigorous the review the less evidence there will be to suggest that the intervention is effective (Petticrew, 2003). He notes that reviews which conclude that good evidence is currently lacking are not always useful to policy and practice. However, it could be suggested that even cautious conclusions are dangerous and should not support recommendations for policy and practice. This appears to be an area of on-going debate, and should be explored in future methodological research.

7.5.4 Consensus?

When asked whether or not they thought there was consensus about the criteria by which the quality of health promotion evaluations should be judged, around two-thirds of those interviewed thought there was probably no consensus. A study conducted by the former Health Development Agency, which asked a range of public health experts to comment on what they thought were the most appropriate types of evidence, reported that there was 'narrow consensus' that RCTs should be used wherever feasible (but acknowledging they may be less feasible for socio-political interventions) (Weightman *et al*, 2005: 8).

It is difficult to reconcile the differing findings of these two studies and draw definitive conclusions about whether there is consensus. This difficulty might be explained by one of the comments made by an interviewee in this study: that health promotion is such a diverse activity it may be unrealistic to expect consensus. 'Forcing an orthodoxy', as described by one interviewee, onto an area with such variability and complexity may never completely be achievable. Consensus may also be unrealistic due to the perception by some of the interviewees that academic groups who routinely produce systematic reviews seldom seem to collaborate with each other on methodological issues. The result is diversity in methods and criteria leading to numerous systematic reviews of similar topics coming to different conclusions. The danger associated with a lack of consensus is that stakeholders will receive mixed messages about the effectiveness and appropriateness of interventions, and this will exacerbate variations in policy and practice, with consequent (potentially negative) impact on health.

There have been some attempts to foster collaboration between systematic reviewers, including the efforts of the Cochrane and Campbell Collaborations (e.g. the Cochrane Health Promotion

and Public Health Field (Jackson *et al,* 2004; Jackson and Waters, 2005; Waters *et al*, 2006)). However, even within the Cochrane Collaboration the various review groups have differing policies on quality assessment, not always cognisant with the standard guidelines on conducing systematic reviews in the Cochrane Handbook (Higgins and Green, 2008). Again, it may not be realistic to expect complete standardisation of methods given that different research groups maybe obligated to use particular methods by their funders or host institutions. Protection of intellectual property may also discourage collaboration if it is felt that academics may no longer have full control of their output. Competitiveness between academic institutions in terms of securing funding and publications may also inhibit alliances (Oakley *et al*, 2005).

Despite these issues there are signs that progress is being made in achieving consensus. It was commented that there is a move towards consensus, or at least a better understanding, on the appraisal of non-randomised studies, based on empirical methodological research done by Jon Deeks and others (as mentioned above, Deeks *et al*, 2003). Generally, it was felt that there was more consensus about how the quality of RCTs might be judged, but less so about non-randomised evidence.

7.5.5 Summary and recommendations

The published systematic reviews of health promotion mapped in this study made favourable comments on the benefits of RCTs and experimental evaluation, although this was balanced by recognition that such methods can be poorly executed, and that in some circumstances observational methods can be useful. There are continued debates in the literature about the appropriateness of including observational studies in the absence of experimental designs, notwithstanding the use of caveats to alert users to potential biases. **Further methodological research should be conducted with a view to reconciling these debates.**

The general perception from those interviewed in this study was that there is little or no consensus in terms of quality assessment criteria, though this might be an unrealistic goal in an area as broad as health promotion. There have been some attempts to develop common models of systematic reviewing in health promotion, although academic competition may inhibit collaboration between different research groups. **It is recommended that further investigation into effective initiatives to foster collaboration and consensus, in order to reduce duplication of effort (to ensure efficient use of public funds) and to lessen the risk of conflicting findings from systematic reviews and the negative knock-on implications for policy and practice.**

7.6 How systematic is quality assessment?

Research objective:

To assess the extent to which quality assessment is conducted and reported in a 'systematic' manner.

- Do systematic reviews of health promotion apply the same set of criteria to each study?
- Do systematic reviews of health promotion single some studies out for criticism over others?
- Do systematic reviews criticise studies for specific methodological flaws without having formally appraised them?

Guidelines for the conduct of systematic reviews encourage thorough assessment of the quality of the evidence (Egger *et al*, 2001; Centre for Reviews and Dissemination, 2009; Higgins and Green, 2008). Yet, half of the reviews in this study were classified as being ambiguous or inconsistent in their approach to quality assessment. There were three overall categories of ambiguity or inconsistency. First, there were reviews that reported using quality assessment criteria but which did not state what the criteria were. Second, there were reviews in which it was not clear whether any criteria were used at all despite the provision of a (sometimes lengthy) critique of the evidence. It appeared that selected studies were critiqued, probably because these were the ones the authors considered to have the most limitations. Whilst it could be argued that highlighting the most serious inadequacies is justified, it is not clear whether there was a systematic process for prioritizing these shortcomings. Third, there were reviews which reported quality assessment criteria but which went on to critique studies on additional issues not covered by the criteria. It is not clear whether each study was systematically assessed in terms of these additional issues.

There are few published methodological studies which have assessed inconsistencies and ambiguities in the assessment of quality. Moja *et al* (2005), who compared quality assessment procedures of Cochrane with non-Cochrane systematic reviews, reported that Cochrane reviews were more likely to state the intention to assess quality (93.7% *v* 63.5%, respectively). However not all of the reviews, Cochrane or otherwise, reported having actually performed an assessment, and about 5% of reviews in each group carried out quality assessment despite not being explicitly stated as an intention in the methods.

The ambiguities and inconsistencies to quality assessment such as those identified in the current study may erroneously influence the findings of a review, particularly if judgement of quality govern which studies are used to support conclusions. The shortcomings of certain selected studies may receive disproportionate attention, whilst limitations of other studies may go unnoticed. As discussed in Chapter 1, the effectiveness of interventions may be over-estimated if bias is present (Jüni *et al,* 2001; Schulz *et al,* 1995). There is the potential danger that the results of a review may over, or in some cases under-estimate the effectiveness of an intervention. It is therefore crucial that systematic reviews are as rigorous and methodical as possible in their assessment of quality in accordance with guidelines. **It is recommended that systematic reviews base their critique of the evidence on explicitly reported quality assessment criteria, so that the basis of their judgement is transparent allowing users of reviews to determine whether or not they are fair and the results of the review are credible.**

7.7 How do systematic reviews use quality judgement?

Research objective:

To assess how systematic reviews of health promotion use quality judgements:

- Do the findings and conclusions of systematic reviews reflect the strengths and weaknesses of the included studies?
- If so, by which methods? (e.g. quality thresholds; quality weighting, etc)
- Is there consensus on the most appropriate method?

As mentioned in Chapter 1, the methodological literature recommends that the strengths and weaknesses of the evidence included in a systematic review are taken into account in the analysis of results and formulation of its conclusions (Detsky *et al*, 1992; Egger *et al,* 2001; Higgins and Green, 2008; Jüni *et al*, 2001; Moher *et al*, 1998). The results of this study show that health promotion systematic reviews mapped in this study have generally observed these recommendations. The vast majority of systematic reviews (just over 90%) that assessed quality took their judgement of the evidence into account in the analysis of effectiveness. This is a higher figure than found in other studies which investigated this issue. The aforementioned study by Moja *et al* (2005) found that just over half of the systematic reviews they surveyed incorporated quality judgement in their interpretation of results. Fewer still were identified by Moher *et al* (1998) who reported that only a quarter of meta-analyses took assessments of quality into account. However, these previous studies included systematic reviews of health care

as well as health promotion, hence they may not be wholly comparable with the current investigation. Nonetheless they help to put the results into perspective.

The reviews mapped in this study were classified as using one or more of four methods of incorporating quality judgement in their analysis. Most of the reviews (81%) used a narrative approach, employing caveats about quality around their conclusions about the effectiveness of interventions. A far smaller proportion (38%) weighted the influence that better quality studies were allowed to exert over the summary of results, and fewer still employed sensitivity or threshold analysis (23% and 15%, respectively).

There is little guidance in the literature as to which method is most appropriate. It could be suggested that a narrative approach is perhaps less valuable than a more empirical method such as sensitivity analysis. The former is characterised by qualitative comments made by reviewers, but with the potential danger that caveats about quality get lost in the overall conclusions (Higgins and Green, 2008). The latter analysis shows how overall effects vary according to the addition or removal of poorer quality studies, and could be interpreted by some as being more objective and informative. This raises the question of whether empirical methods should be prioritised over narrative approaches for considering study quality in the analysis of results in systematic reviews?

A survey of systematic reviewers, methodologists and journal editors conducted by Moher *et al* (1999) found that, in general, sensitivity analysis was the most favoured approach, followed by a threshold approach and then by a narrative approach. The benefit of sensitivity analysis over say, a threshold approach, is that it allows the impact of studies of varying quality to be explored, rather than complete exclusion of poorer studies (which would render the impact of their exclusion on the overall effectiveness of an intervention unknown). Systematic reviews that exclude vast quantities of studies on the grounds of poor methodology have long been criticised by those who consider that the evidence has something to offer (Ogilvie *et al*, 2005; White, 2001), an area of debate discussed earlier (see Section 7.5.3). Indeed, as mentioned in Chapter 1, inclusion of all evidence irrespective of perceived methodological quality is a key feature of alternative approaches to systematic review, such as realist synthesis (Pawson, 2006a; Pawson *et al*, 2005).

The potential downside of sensitivity analysis, however, is that it may generate a range of effectiveness estimates, and these may be misinterpreted by users of reviews unsure of which estimate to trust (Higgins and Green, 2008). Some may even chose the estimate that most favours their particular view point or vested interest, without explicitly acknowledging

methodological weaknesses inherent to the evidence. Further research should investigate the advantages and disadvantages of different approaches.

Two-fifths of the reviews mapped in this study employed a combination of approaches, with the most common being a qualitative discussion of quality alongside weighting of better quality studies. It would seem a sensible approach to use more than one method to ensure that the results and conclusions of a review are as observant of potential methodological biases as possible, in accordance with guidelines (e.g. threshold approach accompanied by sensitivity analysis) (Higgins and Green, 2008).

In general the reviews did not provide a rationale for the approach or combination of approaches used to incorporate quality judgement into their analysis. A similar result was reported by Moja *et al* (2005) who found that only a third of reviews specified how they intended to use their quality judgement. The reporting of a rationale shows that the review authors have given some consideration to the merits and drawbacks of the various methods, and may give the review a greater sense of credibility.

In summary, this study found that the results of the majority of the systematic reviews mapped reflected the strengths and weaknesses of the evidence, mostly using a narrative approach. Only two-fifths of the reviews employed more than one approach. **It is recommended that systematic reviews use a combination of approaches to incorporate quality judgement into the synthesis of evidence, and provide a justification for the methods they have chosen. Further investigation should assess the advantages and disadvantages of different approaches, and where possible empirically test them to examine the impact on the overall results of systematic reviews.**

7.8 The assessment of external validity in systematic reviews

Research objective:

To assess the extent to which systematic reviews of health promotion assess the external validity of included studies:

- For what purpose do systematic reviews of health promotion assess external validity?

7.8.1 To what extent is external validity assessed?

As mentioned in Chapter 1, guidelines on the production of systematic reviews encourage assessment of the external validity of the evidence base (Armstrong *et al*, 2008; Jackson *et al*, 2004; Jackson and Waters, 2005; Centre for Reviews and Dissemination, 2009; Petticrew and Roberts, 2006). It is encouraging, then, that the vast majority of systematic reviews in this investigation (80%) assessed the external validity of included studies.

There was, however, variability in the degree to which it was assessed. Some systematic reviews included assessment of external validity as one of their objectives, whilst others gave it only cursory consideration. Whilst the vast majority of the reviews extracted data on aspects of external validity, relatively few mentioned performing an appraisal of external validity (e.g. using a checklist or instrument). The RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) instrument has recently been devised to enable practitioners, policy makers and researchers to assess external validity of a study and the potential for generalisation (Green and Glasgow, 2006). This appears a promising framework that could be applied within a systematic review to formalise the assessment of external validity (see Recommendations Section (7.8.4) below).

External validity was assessed for three key purposes by the systematic reviews included in this study (Chapter 4). The most common purpose was to facilitate generalisability / replicability (80%), followed by an assessment to explain results (56%), and least commonly, to assess intervention quality (37%). Nearly two-thirds of the reviews were classified as assessing external validity for more than one purpose. The following sub-sections discuss these in more detail.

7.8.2 Generalisability and replicability

As discussed in Chapter 1, the literature suggests that for systematic reviews to be of maximal use to stakeholders it is essential for them to discuss generalisability and replicability (Bonell *et al*, 2006). Accordingly, issues of generalisability and replicability were considered important by the interviewees in this study. Just over two-fifths suggested that systematic reviews should pay more attention to the context within which interventions are delivered, and just under one-fifth suggested particular attention to generalisability and replicability. Their suggestions appear to have been taken up, as the most common aspect of external validity assessed by the systematic reviews mapped in this study was the generalisability / replicability of included studies (80% of reviews).

7.8.2.1 Characteristics of study populations

Within the realm of generalisability and replicability the reviews paid particular attention to the delivery of the intervention (100% of reviews) and the content of the intervention (>90%). The fact that the majority of the reviews (83%) assessed the socio-demographic and other characteristics of the study populations is an encouraging sign. It is important for practitioners and policy makers to judge the extent to which the characteristics of the intervention recipients in an evaluation are reflective of those for whom the intervention is intended in practice (Rychetnik *et al*, 2002). It is generally acknowledged that evaluations, particularly of health care interventions, tend to recruit younger, healthier people more motivated to accept and comply with interventions (Bartlett *et al*, 2005). Likewise, people enrolled in health promotion evaluation studies are often younger and of higher socio-economic status (Britton *et al*, 1998). The 'efficacy' of the intervention in the evaluation setting is therefore often artificially higher than its effectiveness in the real world (Galloe *et al*, 2008; Shadish *et al,* 2000). Health promotion initiatives which seek out those most likely to be receptive to health related behaviour change may, paradoxically, increase health inequalities, as the people who may need the intervention most may be overlooked or may not be in a position to take full advantage of the resources and services provided (Kelly, 2006b; Thomson *et al*, 2008). It is evident then, from the results of the current study, that policy makers and practitioners can use systematic reviews to judge the extent to which effective health promotion interventions are likely to be successful with their own local populations.

Another pertinent issue is that the skills, knowledge and motivation of the intervention providers may differ considerably between the evaluation setting and routine practice (Zhu *et al*, 2002). One of the systematic reviews included in this study noted that the intervention providers were more skilled and motivated than those in practice, and that the intervention recipients, who volunteered to participate, were more likely to benefit than those who had not self-selected to take part (Hillsdon *et al*, 2005). Systematic reviews that neglect this issue may therefore create the false expectation amongst practitioners that their replication of an intervention may achieve similar results.

7.8.2.2 Resources and costs

The feasibility, in terms of resources, of replicating effective interventions was considered by just over two-fifths of the reviews (42%). These reviews commonly discussed the economic resources required for interventions to be successfully replicated, particularly the requirements for interventions evaluated in developed countries to be implemented in resource poor countries

(e.g. Powell *et al*, 2004) (An issue also discussed in detail by one of the interviewees in this study, see Chapter 6).

It is particularly important for systematic reviews to assess resource implications given the increased focus on evaluating the cost-effectiveness of health promotion and public health (Rush *et al*, 2004). The need in the UK for sound evidence of cost-effectiveness was one of the key recommendations of the Wanless report into public health and inequalities (Wanless, 2004). An integral feature of NICE's guidance on health technologies and public health is the assessment of the cost-effectiveness of interventions (Kelly, 2005; Kelly *et al*, 2005; National Institute for Health and Clinical Excellence, 2006). As mentioned earlier in this chapter, there has been increased investment in the UK in health promotion and public health research from the NIHR, and a key requirement of the studies commissioned is the assessment of cost-effectiveness (Allen and Stockley, 2008). For example, in 2008 the NIHR HTA Programme funded a systematic review and economic evaluation of behavioural interventions to prevent sexually transmitted infections in young people (Shepherd *et al*, in press). I was the principal investigator of this project and experienced, first hand, the challenges inherent in assessing the long term costs and outcomes of health promotion activity. Although I have co-ordinated systematic reviews with integral economic evaluations of clinical interventions (Shepherd *et al,* 2006b; 2007b; 2008) this was the first time I had been involved in a systematic review of health promotion that is conducted to inform an economic evaluation. One of the things that myself and the team found was that very few of the primary evaluations reported any form of economic evaluation of their interventions.

Given the absence of primary economic evaluation, detailed information on the resources needed to mount interventions is essential to enable health economists to conduct secondary cost-effectiveness analyses. For example, they need to know the costs of the intervention provider (e.g. salary, overheads), the location (e.g. capital costs of building or hiring the venue) and materials used (e.g. leaflets, videos, computers, etc) (Drummond *et al*, 2005; Øvretveit, 1998). Unfortunately, cost data are not always reported by primary studies, and even basic details on the characteristics of the intervention that would allow health economists to estimate and cost the resources used are often lacking, as was the case in the aforementioned sexually transmitted infections project (Shepherd *et al*, in press). Systematic reviews therefore often have little, if any, comment to make on costs and resources. In such cases it would be desirable for them to acknowledge the paucity of data and at the very least, comment on the likely resource implications of interventions.

The upshot of all this is that systematic reviews of health promotion with integral economic evaluations are likely to become more common to meet policy needs. The methodology for these reviews therefore needs to advance, and investments will be needed to recruit and train people with the necessary skills so that there is adequate capacity to meet the demand.

7.8.3 Quality of the intervention

In this study the aspect of external validity least commonly assessed by the systematic reviews was the quality of the health promotion intervention, as classified in just over a third of reviews. This finding is perhaps not surprising as within evidence-based health promotion the quality of interventions appears to have received comparatively less attention compared to the quality of evaluation, despite an evolving literature on quality assurance (Catford, 1993; Evans *et al*, 1994; Speller *et al,* 1997, Tones 2000; van Driel and Keijsers, 1997).

This appears, then, to be an area in need of further methodological development. Initiatives are currently taking place which may, in part, facilitate progress. For example, in June 2008 the International Union for Health Promotion and Education (IUHPE) and the Society for Public Health Education (SOPHE) organized an international consensus conference to agree the domains of core competency in health promotion and health education. The output was a consensus statement covering eight domains of core competency necessary for effective health promotion practice (e.g. planning, implementation, evaluation). This work is perhaps at a broader, more general, level than necessary for the assessment of quality within a systematic review, which would be concerned with more context-specific issues such as the appropriateness of an intervention within its intended setting. Nonetheless, it endorses the importance of quality assurance, and provides a framework for future developments.

Herbert and Bø (2005) recommend that systematic reviews should routinely assess the quality of the intervention, and to incorporate this into their analyses to help explain the effects observed (e.g. by assessing how the effectiveness varies according to presence or absence of interventions judged to be of poor quality). They acknowledge, however, that the criteria by which the quality of an intervention should be judged is one of the most difficult methodological issues facing systematic reviews.

It is beyond the scope of this study to propose exhaustive criteria by which the quality of an intervention can be judged. However, it is possible to explore some of the potential markers of quality which emerged from the results of this study, in terms of case studies. The markers were

heterogeneous, and included use of theory, intervention implementation, and ethics. These are discussed, in turn, in the following sub-sections.

7.8.3.1 Theory

As discussed in Chapter 1, the rationale for using theory to underpin health promotion has been extensively articulated in the literature (Fisher and Fisher, 2000; Green, 2000; Rothman *et al*, 2004; Tones and Tilford *et al*, 2001; Turner and Shepherd, 1999; Wight *et al*, 1998). A central feature of realist synthesis, for example, is the conceptualisation of the mechanism through which the intervention might work (Pawson, 2006a; Pawson *et al*, 2005). This conceptualisation is extended and refined as literature is identified, assimilated and appraised. Whilst none of the systematic reviews included in this study considered themselves realist syntheses (and therefore do not necessarily use theory as their starting point), only just over a third of them reported whether or not their included studies were theory-based. This is surprising given the importance attached to theory in the scientific literature. Furthermore, only two made explicit reference to theory as a marker of quality. In the remainder of the reviews theory appeared to be important in the wider discussion about the evidence, but it was not explicitly mentioned why.

It is commonly assumed that theory-based interventions are more likely to be effective than those with no theoretical basis (Nutbeam and Harris, 2004). Intuitively this makes sense, and would be hard to disagree with, but experimental evidence to support this assumption is lacking. Although the role of theory as a mediator (amongst other mediators) of intervention effectiveness was examined by some of the reviews included in this study qualitatively (e.g. Ciliska *et al,* 2000; Kirby *et al,* 2006), few empirical studies have specifically tested the association between use of theory and effects (Bonell and Imrie, 2001). It is acknowledged, however, that such testing would be difficult, particularly to isolate the influence of theory from competing contextual factors in the evaluation setting.

Nonetheless, guidelines for systematic reviewing recommend that reviews use theory to explain the effects of interventions, although it is acknowledged this can be a challenge as primary evaluations do not always report whether or not theory was used or why it was chosen (Bonell and Imrie, 2001; Jackson and Waters, 2005). This may be redressed by the aforementioned TREND statement, which requires evaluation studies to state which theory was used to guide an intervention (Des Jarlais *et al*, 2004).

A recommendation from this study would be for systematic reviews to discuss the mechanism through which the intervention in question is thought to operate, with reference to all relevant

theories. They should routinely assess the use of, and justification for, theory in their included studies, and discuss whether they consider the theories used were appropriate to the aims of the interventions. Where relevant and feasible, they should also assess the association between theory and intervention effects (e.g. using statistical regression analysis).

7.8.3.2 Fidelity of intervention implementation

The implementation of the intervention was mentioned as an important issue by two of the interviewees in this study, but was formally assessed by only one of the reviews. The reason why few of the reviews assessed implementation might be due to limited detail reported by primary studies. One of the interviewees in this study commented that primary evaluation studies often fail to report adequate data, and systematic reviews therefore tend to neglect discussing the development and delivery of the intervention.

Empirical studies support this assertion. For example, Dane and Schneider (1998) found that only 39 of 162 (24%) evaluations they surveyed reported procedures for the documentation of fidelity. Similarly, Bonell *et al* (2006), in an analysis of eight evaluations of HIV prevention from a systematic review (Rees *et al*, 2004) (a review which I participated in, see the 'Introduction' to this thesis), found that six of them reported integral process evaluation, but only three collected data on the planning, delivery and receipt of the intervention. It is hoped that initiatives such as the extension of the CONSORT statement to improve reporting of RCTs of non-pharmacologic interventions (Boutron *et al*, 2008), and the TREND Statement for non-randomised studies (Des Jarlais *et al*, 2004) will raise reporting standards. As recommended earlier in Section 7.3, academic journals which publish articles on health promotion should sign up to these statements, and evaluation should monitor the effectiveness of the statements in raising reporting standards, and in turn whether systematic reviews pay more attention to these issues.

7.8.3.3 Ethics

Study ethics was not discussed in any great detail by the interviewees, and was not formally assessed by the systematic reviews. However, on occasion the reviews commented on interventions they considered to be ethically questionable, such as the breastfeeding example from the review by Dyson *et al* (2005) (see Chapter 4, Section 4.7.3). This raises an important question, should systematic reviewers govern whether or not a study can be included in a review on grounds of ethics? Does this amount to censorship? Given that ethics is to some extent a subjective concept should it not be left to the reader to judge for themselves whether or not a study is acceptable? It is worth noting that a systematic review of the reporting of ethics in

clinical trials found that the standard of reporting was higher in better quality trials (Ruiz-Canela *et al*, 2001). The fact that poorer methodological quality studies are often omitted from systematic reviews may mean that, by default, those remaining will be ethically more acceptable, although in the absence of further evidence this is only a tentative assumption, requiring empirical investigation.

Despite calls for ethics to be routinely assessed in systematic reviews (Weingarten *et al*, 2004) there does not appear to be much guidance to systematic reviewers how to handle this issue, or consensus on what an ethical intervention might be. The Cochrane Collaboration's guidelines for systematic reviews of health promotion and public health briefly discusses ethics in relation to health inequalities, and encourages reviews to assess the effectiveness of interventions for people of different socio-economic status (Armstrong *et al*, 2008). Such an assessment will identify whether the needs of those at the most disadvantage are being met, and is in accord with current UK health policy with its focus on health inequalities (Department of Health, 2004). However, other than inequalities there is no mention of other aspects of ethics that might arise during a systematic review.

What other ethical issues should reviews therefore assess? Weingarten *et al* (2004) proposes a number of ethical considerations for the assessment of clinical trials, including declaration of financial support, conflict of interests, justification for use of placebo drugs, informed consent, risk of adverse events, and the protection of clinical and personal data. In terms of health promotion, Harden and Oliver (2001) examined the ethical qualities of studies included in EPPI-Centre systematic reviews for young people. This was based on data routinely conducted in two of their systematic reviews. The context for their assessment was the drive over recent years to involve lay people in decision making about health services, observing the principles of empowerment and participation from in the Ottawa Charter. They collected data on whether the interventions were based on needs assessment and developed with input from its intended recipients, whether the recipients were consulted on the evaluation design, whether informed consent was sought, and whether the findings of the evaluation were fed-back to the participants. This is one of the few empirical studies of the ethical dimensions of health promotion within the context of health promotion, and provides a useful starting point for future work. However, further research is required to identify which ethical aspects of interventions that should be prioritised (as there could be many), to give systematic reviews a firmer foundation upon which to assess ethics, where appropriate.

7.8.4 Summary and recommendations

This study found that the majority of systematic reviews assessed external validity, in accordance with guidelines on the production of systematic reviews. However, the extent to which it was assessed varied. **It is recommended that systematic reviews use a formal instrument for the assessment of external validity, such as the RE-AIM framework which appears to have promise as a mechanism for determining generalisability of research findings.**

Generalisability and replicability were commonly considered, in terms of details of the intervention delivery and content, and the characteristics of the study populations. Less commonly assessed by the systematic reviews was the economic cost of providing interventions. **Given the increased focus on demonstrating cost-effectiveness in policy making it is recommended that systematic reviews of health promotion routinely assess the costs and resources of health promotion. More investment will be required to ensure there is adequate capacity, in terms of systematic reviewers with at least a basic understanding of health economics, to meet the demand for evidence of cost-effectiveness. Primary evaluations of health promotion interventions should be accompanied by economic evaluations where possible, or at the very least they should clearly report the resources used to provide the interventions to allow a secondary assessment of cost-effectiveness.**

The study also found that the quality of health promotion interventions was assessed by comparatively few systematic reviews, despite calls in the literature for greater attention to quality. Some of the potential markers of intervention quality were explored in this thesis, notably the theoretical basis of interventions and study ethics. **It is recommended that systematic reviews assess the appropriateness of theories cited by studies, to explain the mechanism through which interventions are purported to work. Guidance on the most appropriate markers of ethical soundness is needed to assist reviewers in handling studies which may be ethically dubious. Further research is needed to identify other markers of intervention quality in health promotion which could be employed by systematic reviews. Consensus development methods involving a range of stakeholders might be a useful way to prioritise quality markers.**

An over-arching recommendation is that primary evaluations should report as much detail as possible on all aspects of the intervention and its delivery to enable systematic

reviews to make a thorough assessment of external validity, whether generalisability, replicability, fidelity or intervention quality. The extended CONSORT statement and the TREND statement on reporting standards should be adhered to by academic journals in the area of health promotion. The effectiveness of these statements and other initiatives to improve reporting of evaluation should be monitored.

7.9 Who conducts systematic reviews of health promotion, and why?

Research objective:

To assess which types of people commonly participate in the production of systematic reviews of health promotion:

- Who does reviews (e.g. academics, health and other professionals, lay people), and what is their rationale for doing them?
- Who performs quality assessment in systematic reviews? (e.g. people who specialise in producing systematic reviews; people who specialise in the topic area being reviewed; combinations of these)
- To what extent are systematic reviews the product of collaborative teams? What are the advantages and disadvantages of collaborative team working?

7.9.1 Who is involved in systematic reviews of health promotion?

As found in the mapping exercise (Chapter 4), the most common type of person involved in the production of the systematic reviews sampled in this study were those with an academic affiliation (two-thirds of the reviews). Practitioners were also involved, but to a lesser extent (one fifth of the reviews). However, it is likely that some of those with an academic affiliation may have also held professional appointments, and vice versa, as was the case with a two-fifths of those interviewed in this study (Chapter 6). This finding is acknowledged by Hammersley (2006) who commented that reviews may be produced by people wearing both academic and practitioner 'hats'.

Around a fifth of reviews were classified as being produced by both researchers and practitioners. That is, there was at least one author with an academic affiliation and one with a professional affiliation. Other types of people (lay people, policy specialists) were virtually absent from the authorship of the reviews, although it is possible that they may have been consulted as stakeholders in an advisory category. However, use of an advisory group was only

reported by two of the reviews. Although there was some evidence of multi-disciplinary collaboration (see Section 7.9.3 below), the results of this study suggest that systematic reviewing is still very much the preserve of academia. This is at odds with the 'Best Research for Best Health' policy, and the assertion by at least one of the interviewees in this study, that it is desirable to increase participation from all stakeholders in the production of systematic reviews (Department of Health, 2006).

7.9.1.1 How do we involve stakeholders in systematic reviews of health promotion?

'Best Research for Best Health' does not make specific reference to how health professionals (or any other stakeholders for that matter) might be encouraged to contribute to systematic reviews. The question arises: What might be the barriers or facilitators to involving stakeholders in systematic reviewing? As discussed earlier, the considerable amount of time and resources necessary to conduct systematic reviews, particularly of complex topics such as health promotion, might be one explanation. Health promotion practitioners may find it difficult to secure dedicated time and resources to participate. Their employers, including Primary Care Trusts and non-statutory agencies, may not be in a position to fund training and grant study leave given the immense pressure on their resources and budgets. Practitioners may also potentially perceive systematic reviewing to be an academic endeavour requiring skills beyond their capabilities, an issue discussed in more detail later (Section 7.11.2).

The barriers to involving the public in systematic reviewing may be similar to those facing professionals, such as lack of time, a particular issue to those in full time employment or with family commitments. They may also be dissuaded by the perception that specialist knowledge and skills are needed. However, it is often personal experience of a health condition that encourages people to advocate on behalf of others in their position. Indeed, their personal perspective may be actively sought by systematic reviewers wishing to make their work more clinically meaningful. Whether they have, or can develop, the skills necessary is less certain and is discussed below (Section 7.11.2).

Further research is therefore needed to assess how all stakeholders can be involved in the production of systematic reviews in health promotion. This might take the form of further qualitative research to investigate the issues arising from this project, followed by evaluation of promising strategies.

7.9.2 The role of topic and systematic review specialists

The mapping of systematic reviews in this study also found that reviews were predominantly conducted by people classed as being specialists in the topic under review (around two-thirds of the reviews). Furthermore, just under half of those interviewed said that they had become involved in systematic reviewing because of their expert knowledge of the topic area, or because they specifically wanted to advance knowledge in their area. These findings suggest Hammersley's argument, that the critical assessment of research evidence cannot be reduced to a set of procedures without substantive topic knowledge and broader judgement, may be less of an issue (Hammersley, 2006). Nevertheless, whether these reviews are of a higher standard (however defined and measured) compared to those lacking in expert input, however, is a matter of debate. There have been very few published attempts of comparisons between expert and non-expert led systematic reviews. Meads (2007) undertook a small-scale study analysing systematic reviews conducted by students for their dissertation as part of the MSc in HTA at the University of Birmingham. Students classified as having a clinical background in the topic under review had slightly lower marks than those without such a background (no statistically significant difference). This was an exploratory piece of work and is an area in need of further research, although it may prove difficult to formally evaluate without use of subjective markers of what a high quality systematic review is.

A smaller number of reviews were conducted by people classified as specialists in systematic reviewing (just under a quarter). Likewise, just over a third of those interviewed in this study said that they had been recruited to a job specifically to conduct a systematic review, although not all of them remained in that job at the time of the interview. There was some evidence to suggest a blurring of the dichotomy between topic specialist and systematic review specialist. For example, as mentioned in Chapter 4 (Section 4.6.2), one of the authors of a systematic review included in the mapping exercise (a topic specialist in child and maternal nutrition) had become increasingly involved in systematically reviewing topics in her specialist area. This cross-over in expertise also works in the other direction. Those whose job is to routinely conduct systematic reviews (such as myself) often have to review a range of topics which they may not familiar with due to the changing needs of policy makers. Invariably this necessitates a rapid learning curve each time a new review is initiated. The one consistency these reviews have is their familiarity with the methodology, but some may also have the opportunity to develop expertise in a particular topic area.

For example, I have been fortunate to build up some degree of knowledge of the treatment of chronic hepatitis. My department, who have a contract with the Department of Health to

conduct technology assessment reports for NICE, was commissioned to produce a systematic review of the clinical and cost-effectiveness of pharmacological treatments for moderate to severe chronic hepatitis C, which I co-ordinated (Shepherd *et al*, 2000). We were requested to update this a few years later to incorporate the launch of new drugs for the condition (Shepherd *et al*, 2004), and then again more recently to assess include patients with the milder form of the disease (Shepherd *et al*, 2007b). On the back of our track record in this area we were also commissioned to review treatments for chronic hepatitis B (again for NICE) (Shepherd *et al*, 2006b), and, again, to update this (Jones *et al*, 2009). Our progress has benefited from a long-term stable funding grant and such opportunities may not be available to all, like us, whose primary role is to conduct systematic reviews.

The result of our specialising in hepatology, I would like to think, is a better quality and more clinically meaningful systematic review, benefiting from expertise in both the methodology and the topic area. But there are other ways to achieve this. One obvious approach is for collaboration between experts in the topic under review and those conversant in the production of systematic reviews, as recommended by Cochrane guidelines (Higgins and Green, 2008). One-fifth of the reviews mapped in this investigation were classified as being produced by both topic and systematic review specialists. Given that around a quarter of the reviews were classified as being conducted by systematic review specialists this shows, therefore, that they nearly always collaborate with 'experts' in the topic area. Again, this casts doubt on Hammersley's concerns (Hammersley, 2006).

7.9.3 Collaboration and team working

It emerged from the interviewees that sometimes teams are convened specifically to do a particular review, and sometimes they are done by pre-existing teams of systematic reviewers (such as myself) as part of a funded programme of work (e.g. for NICE or the Department of Health). As mentioned by one of the interviewees in the latter camp, it can be challenging reviewing topics outside of one's area of expertise, but the complimentary backgrounds of team members help. This finding illustrates that academics who become systematic review specialists in health promotion are drawn from diverse backgrounds. In her opinion the team benefits from collective knowledge of the social, economic, demographic, biological, psychological aspects of health, effectively ensuring that all bases are covered. However, in her experience it is still necessary to involve relevant stakeholders in an advisory capacity. She commented on the advantages and disadvantages of her experience of using advisory groups, but there appear to be few published accounts in which systematic reviewers reflect on or evaluate their experiences

(Rees *et al*, 2004). Further investigation would therefore be useful to help improve the way in which such groups work in the future.

As reported in Chapter 6, the interviewees discussed the advantages and disadvantages of collaborating team working. For example, it was suggested that a pre-requisite for effective quality assessment is open-mindedness and suspension of personal / professional assumptions for the sake of scientific debate. People who routinely do systematic reviews may be more aware of this than, perhaps, those drafted in to participate in a review for their expertise in a topic area.

It was also commented that, although a multi-disciplinary team benefits from the specialist knowledge that each team member contributes, it can be difficult to reconcile their diverse perspectives and research interests in setting a topic for review. However, not all accounts of multi-disciplinary team working have encountered such difficulties. For example, Nind (2006) reported that the experience of agreeing a research question and identifying research priorities within her team was largely positive.

7.9.3.1 Advantages and disadvantages of team working

Whilst multi-disciplinary teams might bring a variety of perspectives to the table this strength may also be a weakness. Tasks such as quality assessment require a shared understanding of the criteria and how they should be applied in order to ensure consistency (Higgins and Green, 2008) although as discussed, Hammersley (2006) would argue that quality assessment cannot be reduced to just a set or rules and procedures. The more people undertaking this task the greater the potential for the inevitable differences in interpretation and subjective judgement, described earlier (Section 7.1.1). Paradoxically, a high degree of consensus and inter-rater reliability by established teams was mentioned by one interviewee as problematic. It was suggested that potential study biases might be overlooked as the team are 'tuned into' the same issues and may not spot things that 'fresh eyes' might see. It is for this reason that the interviewee in question mentioned that she routinely invites other stakeholders to participate in their reviews (e.g. public health doctors). The findings of this study endorse the ideology that systematic reviewing should be a product of collaborative team work, to keep the potential biases associated with both systematic review specialists and topic specialists in check.

7.9.3.2 Practicalities of team working

Whilst collaborative systematic reviewing is favoured by research funders, the findings of this study show that there are practical difficulties in seeking collaborators. As one interviewee

noted, the ability to convene a team is dependent on the availability of suitable people in a position to collaborate. It cannot, therefore, be guaranteed that likely collaborators will always be in a position to commit themselves. Again, the issue of a lack of time may be a barrier. If this ideology is to be successful in practice then adequate funding will be necessary to free-up the time of those who can make a useful contribution. As discussed earlier, whilst there is genuine altruism for evidence based health the days of reviews done on a 'shoe-string' budget may be long gone (Section 7.2.1).

7.9.4 Summary and recommendations

This study has found that systematic reviews of health promotion are predominantly conducted by academics with expert knowledge of the topic under review. The desire that reviews be the product of multidisciplinary collaboration by all representatives of all relevant stakeholders does not appear to be a reality. **It is recommended that further research is conducted to assess the barriers and facilitators to involving all stakeholders in the production of systematic reviews, to inform evaluation of promising approaches to seeking stakeholder involvement.**

There was some evidence of collaborative team work in the production of systematic reviews, predominantly involving academics and health professionals. Whilst this research has identified some of the drawbacks associated with team working, overall it seems that these are outweighed by the benefits. **A recommendation from this research is that systematic reviews should be conducted by multi-disciplinary teams, where possible, involving all relevant stakeholders. Teams should be carefully monitored and evaluated to ensure efficiency and fairness in the judgement of evidence.**

7.10 How do people learn to do systematic reviews?

Research objective:

How do people learn to do systematic reviews of health promotion?

- Which learning strategies are considered most successful?
- What are the barriers, to and facilitators of, learning?
- What are people's experiences of receiving training?

The interviewees were asked to describe and discuss their experiences of learning to do systematic reviews. Common ways of learning included: practical experience of doing reviews, and support from colleagues, mentors, and training courses, each of which were mentioned by around two-thirds of those interviewed. Literature and written resources were also mentioned, by just over half.

7.10.1 Learning through practice and from colleagues

In general the interviewees learned through a combination of methods with a strong emphasis on practice-based learning, as typified by the comment from one interviewee: 'But 99% of it has been on the job training and self-learning'. That this should be such a significant learning strategy sounds intuitive. After all, few people would probably disagree that one of the best ways to learn a new skill is to practice it. It shows that learning systematic reviewing is not radically different to mastering other research methods.

The interviewees also tended to learn from working with more experienced systematic reviewers, and learning within the context of a team emerged as particularly beneficial. In contrast, a handful of interviewees mentioned that their first involvement in systematic reviewing was during the mid 1990s when there were fewer people who they could consult for advice. Much of their learning, therefore, was through trial and error. This point was eloquently illustrated by one interviewee who remarked that it was 'the blind leading the blind' (Chapter 6, Section 6.7.2). Today, it would appear, there is a larger pool of people with skills in systematic reviewing from whom guidance and mentorship can be sought. This increase in capacity has almost certainly developed as a result of increased funding of systematic reviews (as discussed earlier, in Section 7.1), which, in turn, has created greater demand for skilled reviewers. The more people that learn to do reviews the more they can pass on their knowledge and skills to others, a phenomenon which could be likened to professional peer education.

7.10.2 Training

The extent to which the interviewees participated in training varied, from courses lasting one day to those over several weeks. However, the timing of training was not always what might be expected. For example, some only did their training after they had practical experience of reviewing, although there did not appear to be a particular reason for this. It might be that some people prefer hands-on experience in the first instance, and to then consolidate and contextualize their learning through training. They may benefit more from training if they have some personal experiences to draw on, rather than being taught about something that, to them, is largely abstract.

Access to training did not appear to be a barrier, and this is reflective of the investment that has gone into widening access to training and increasing capacity to meet the demand for reviews. For example, in the UK the ESRC established the National Centre for Research Methodology (NCRM) in 2004 and one of its 'nodes' specializes in methods for evidence synthesis, within the social sciences. One of this node's key achievements has been to provide training courses on systematic reviews in the UK (Wiles and Bardsley, 2008). The NIHR has also invested through its Research Capacity Development Programme (RCD) which runs a fellowship scheme for 'Research Scientists in Evidence Synthesis' awards (National Institute for Health Research, 2008). The awards are more geared towards supporting a smaller number of promising academics in advancing the methodology of evidence synthesis, as opposed to providing training to a wider audience. Nonetheless, these examples show that building capacity for systematic reviewing is firmly on the agenda, and that opportunities for training and development are available.

However, the ability to take up such opportunities requires adequate time and funding, as commented by the interviewees. One of the interviewees, from a resource poor country, commented that he probably would not have had the opportunity to receive training had he not received a scholarship from the Cochrane Collaboration who funded him to visit the UK. The Cochrane Collaboration, and in particular the Cochrane Health Promotion and Public Health Field, is committed to reducing global health inequalities and encourages people in resource poor countries to participate in the production of systematic reviews (Doyle *et al*, 2005). For this commitment to remain a reality there needs to be continued investment in training and support. The 'Reviews for Africa' programme, mentioned in Chapter 6, is an example of such commitment.

The general consensus seemed to be the training received had been adequate, although again, it was mentioned that the extent to it can compensate for practical experience is limited. Numerous suggestions were made by the interviewees for improving training, including use of practical exercises, provision of longer training courses, more coverage of complex interventions and non-experimental evaluation designs, and greater use of mentors. Some of these issues are expanded upon below in Section 7.11.

7.10.3 Literature and written resources

Just over half the interviewees cited literature as being helpful, particularly as a resource to consult in conjunction with other forms of learning. In parallel with the increase in capacity for

systematic reviews, in the last few years there has been a proliferation of guidelines and text books in the area of evidence-based health care (Egger *et al*, 2001; Glasziou *et al*, 2001; Higgins and Green, 2008; Khan *et al*, 2003; Centre for Reviews and Dissemination, 2009; Torgerson, 2003), evidence-based nursing (Cullum *et al*, 2008; Dicenso *et al*, 2004) and social sciences and social care (Coren and Fisher; 2006; Petticrew and Roberts, 2006).

However, there was some suggestion by the interviewees that currently available guidelines do not adequately address issues relating to systematically reviewing complex interventions. Since the interviews were conducted, the Cochrane guidelines for health promotion and public health (assembled by an international Taskforce of which I was a member) have been included as a chapter in the Cochrane Handbook for reviewers (Armstrong *et al*, 2008). As the Handbook is one of the key resources for systematic reviewers within and outside of the Cochrane Collaboration, the guidelines are now more accessible than before. It is unclear, however, whether there are any plans to evaluate the guidelines. Continued monitoring and revision of these resources is therefore essential to ensure they remain relevant and reflect methodological developments in the field as they happen.

7.10.4 Summary and recommendations

In this study it was found that systematic reviewers tend to develop their knowledge and skills through a combination of learning approaches, a key strategy being hands-on practical experience often with more experienced reviewers. Over the years it appears that capacity has increased and there are more skilled reviewers in a position to pass on their learning to others. **It is recommended that people wishing to learn to conduct systematic reviews be given the opportunity to have hands on experience, in conjunction with other learning strategies as appropriate.**

There appears to be more commitment to providing training for systematic reviewers, yet reviewers may not always be in a position to take advantage of them due to factors such as geographical location, and availability of time and funds. **It is recommended that commitment to training is extended in terms of provision of bursaries and scholarships to increase access to courses. This should be done in conjunction with initiatives to increase wider stakeholder participation in systematic reviews, as also recommended in Section 7.9.**

Guidelines and key texts on systematic reviewing were considered by some of interviewees in this study as being a useful reference. These guidelines have an important role to play in terms of setting standards for systematic reviewing. **It is therefore essential that they are updated**

regularly to take into account new developments in methodological research, to provide systematic reviewers with clarity around issues such as quality assessment, as also recommended in Section 7.4.

7.11 Helping others to learn to do systematic reviews

Research objective:

What are reviewers' experiences of helping others to learn systematic reviewing?

- What forms of training and support are provided?
- What issues and topics are covered?
- What have been the challenges and successes in providing training and support?

7.11.1 Key characteristics of training provided

The interviewees in this study described their experiences as providers of training and support (Chapter 6). The majority of them (76%) reported some experience as providers. Their involvement varied from full time training officer, to informal mentor. Nearly three quarters had provided training to professionals (mostly health professionals), two-thirds had taught on an academic degree course (generally to post-graduates, and mostly as one component of syllabus), and around a quarter had provided training for Cochrane, or had been mentors.

The training tended to cover most of the stages of a systematic review, with variations in terms of length and level of detail. Where training was provided as part of a degree course, students were required to undertake a systematic review as part of the course assessment (e.g. a dissertation). The philosophy behind this is that an element of practice-based learning within the context of an academic course is an effective way to put learning into context.

Evidence-based health appears to be a standard topic in a growing number of a degree and diploma courses in health (Douw *et al*, 2002). In the UK it is included as part of the syllabus for nurse training and the medical curriculum (General Medical Council; 2003; Parkes *et al*, 2001), and at higher degree level it features in some Masters of Public Health courses, such as that run by the London School of Hygiene and Tropical Medicine. There are also Masters degree courses specializing in various aspects of evidence-based health, for example the MSc in Evidence Based Health-Care run by the University of Oxford; the MSc in Evidence for Public

Policy and Practice run by the EPPI-Centre; and the MSc in HTA, offered by the University of Birmingham (Taylor *et al*, 2002).

The existence of these courses suggests the legitimisation of evidence-based health in higher education, and will likely cultivate a new generation of health professionals with some degree of awareness and competence in the principles and practice of systematic reviewing. It will go some way to increase participation in systematic reviews by practitioners, notwithstanding the practical barriers mentioned earlier. It is unclear, however, if these courses adequately cover issues relevant to health promotion, or whether degree courses in health promotion (mostly available at postgraduate level) adequately cover evidence-based health. A more detailed analysis of the syllabi of current higher degrees of health promotion is necessary.

7.11.2 Challenges in providing training

The interviewees identified two common issues that, in their experience as trainers, trainees tended to find difficult to comprehend. The recurring issues of subjective judgement, and pressure on time and funds were at the root of these challenges.

7.11.2.1 Statistics

The first issue identified was a difficulty in understanding the statistical techniques involved in systematic reviewing, as mentioned by just over half the interviewees. This is not necessarily a problem unique to health promotion, but is exacerbated by the fact that (mentioned earlier, Section 7.2.2) many health promotion evaluations allocate clusters rather than individuals to study groups, to reduce the likelihood that the people in the comparison group might receive the intervention (Torgerson, 2001). An evaluation of a classroom-based peer-led school sex education intervention which randomised schools to intervention or comparator is a typical example of cluster trial (Stephenson *et al*, 2004). However, cluster allocation requires different statistical assumptions to those commonly used in meta-analyses of studies in which individuals are the unit of allocation (Killip *et al*, 2004; White and Thomas, 2005). This makes systematic reviewing health promotion a more complex task, and places greater pressure on reviewers.

The interviews also reported difficulties with the interpretation of statistical results. It was commented that the results of statistical tests used in systematic reviews can often be open to interpretation, and that trainees find it difficult to cope with this uncertainty. The need to sometimes exercise judgement was mentioned as a being a difficult message to communicate to novice reviewers lacking in confidence to do anything other than follow explicit instructions.

Difficulties in comprehending statistical issues was mentioned as being a particular problem in shorter training courses, where the limited time available is reserved for introducing basic concepts of reviewing, rather than discussing the finer detail. This suggests that longer training is required to adequately teach statistical issues, but this has obvious implications for time and costs as discussed earlier. It also raises the question of how realistic is it to expect all reviewers to be statistically numerate? If systematic reviews tend to be produced by multi-disciplinary teams, as the results of this study show, then it may not be an essential pre-requisite for all reviewers to be skilled in statistics as long as there is some expertise within the team.

7.11.2.2 Quality assessment

The second difficulty mentioned by the interviewees was in understanding the rationale for, and methods of, assessing quality. This seemed to be more of a problem for those without an academic background. For example, it was commented that trainees, particularly those who have returned to studying from practice, are often unaware of the need to think critically, and liable to accept evidence on face value. Those with some grounding in research tended to do better at exercises to develop critical appraisal skills than those without. Does this suggest, then, that a pre-requisite for systematic reviewing is possession of a higher education qualification?

Hammersley (2006: 248) makes a distinction between the capabilities of academics and professionals, suggesting that a universal competency is inappropriate:

"We must not pretend that policy-makers and practitioners can or ought to operate in exactly the same manner that is required of researchers. Policy–making and professional practice do not call for the same cognitive or ethical orientation as research…they cannot do this in the same way as researchers"

There is little published literature on core competencies for systematic reviewing. Most academic posts for systematic reviewers, including those in my department at the University of Southampton, require applicants to hold a first degree, preferably with some post-graduate experience. Other institutions, such as the CRD at the University of York require a Masters level degree as an essential pre-requisite. One of the interviewees in this study mentioned that although most team members are given the opportunity to participate in critical appraisal, there is a perceived minimum level of expertise necessary. He commented that he probably would not assign the task to an under-graduate, which is in contrast to comments from other interviewees who said that, in their experience, medical and nursing students tended to grasp the principles relatively easily.

If some notional threshold of ability is necessary to facilitate quality assessment in a systematic review then an even higher level of expertise would be necessary to engage in a realist form of synthesis (Pawson, 2006a; Pawson *et al*, 2005). The complex process of theorising the complexity of the intervention, gathering and assimilating data, revising the theory, assessing the relevance and rigour of the evidence and coming to conclusions and recommendations would likely be a challenge for many experienced academics, as acknowledged by Pawson himself. This casts doubt upon the likelihood of this alternative, and promising, approach to evidence synthesis developing itself and making its mark in the policy arena.

All of the preceding discussion is at odds with the ideology, as discussed earlier, of involving non-academic stakeholders in the production of systematic reviews. It is also contra to the finding from this research that people tend to learn most through actually producing reviews. If reviewers are not considered to be competent to participate in tasks such as quality assessment they are being denied the opportunity to learn.

There is some evidence that the needs of those who train via a non-academic route are addressed. One interviewee commented that the MSc course she co-ordinates is designed to encourage students to apply the skills they use to appraise phenomena in everyday life to evaluate research evidence. The idea is to emphasise that critical appraisal is a skill that we all have, and that this will therefore give trainees greater confidence to develop them in the context of evidence-based health. She did not comment on how successful this had been but it does sound like an appealing approach, and one that might also potentially be effective with lay people.

As noted in Chapter 1 the evidence for the effectiveness of teaching systematic reviews skills to non-academics is relatively small. The findings of earlier studies have been overtaken by changes in practice and advancements in methodology discussed earlier in this chapter (Milne and Oliver, 1996; Oliver and Peersman, 2001). Further investigation into the training needs of health promotion practitioners and other stakeholders, and the effectiveness and appropriateness of different approaches to teaching them critical appraisal skills, including that mentioned above, would be welcome.

Picking up on an issue mentioned previously, one of the interviewees commented that some of her trainees perceive that critical appraisal is a task that does not require judgement. They practice appraising studies using highly structured instruments, the benefit of which is to guide them through the process, reassuring them and building their confidence. The downside is that

they end up following what she called 'the cookbook'. The danger is that they do not think about the wider implications of their work and consider what it really means. This problem, she felt, was particularly evident among medical students, who operate in an environment dominated by procedure. As commented throughout this chapter, some degree of subjectivity is inevitable particularly in judging quality. It can be daunting practising critical appraisal without an explicit set of procedures and criteria to rely on. To be encouraged to deviate from this, where necessary, may be disconcerting. The challenge, it would seem, is to communicate this to trainees without undermining their confidence. This is an issue that should be investigated further.

7.11.3 Summary and recommendations

This study has elicited the perspectives and experiences of those who provide training on systematic reviewing. Particular issues that have been found to be challenging to teach include the statistics used in meta-analysis, and the process of quality assessment. In both cases the potential for differences in interpretation and the need to exercise judgement are difficult issues for trainees to grasp. Some of the strategies that trainers have used to overcome these challenges have been discussed. **Further investigation is needed to identify potentially effective ways of addressing these issues within the context of training. Potentially effective strategies could be implemented and evaluated to assess their impact.**

The findings of this study raise questions about the level of competency needed to conduct systematic reviews, and whether it is realistic to expect reviewers to be skilled in all of the diverse tasks necessary to produce a review. **These findings lend further support to the earlier recommendation (Section 7.9) that where possible systematic reviews be conducted by multi-disciplinary teams, notwithstanding the challenges noted in securing stakeholder involvement.**

7.12 Chapter summary

This chapter has thoroughly discussed the findings of both stages of this research, taking each of the research objectives in turn and examining the implications for the field. The next chapter discusses the strengths and weaknesses of the study, reflecting on the methods used, and my role as a researcher in this field.

Chapter 8 - Strengths and limitations of this study

Chapter outline

The aim of this chapter is to reflect on the methods used to conduct and report this research and to briefly discuss its strengths and weaknesses. In doing this I pay particular attention to my role as both investigator and systematic reviewer, and how this may have influenced both the findings of this research, and my own perspectives. I also critically discuss the overall methodological framework for the research, the scope of the study, and the internal and external validity of the data collected.

8.1 The role of the author

This study has been conducted by someone who routinely conducts systematic reviews of both health care and health promotion. Despite a belief in the value of reviews (and it would be difficult not to do my job without some faith in their contribution to decision making!) I have questioned their strengths and weaknesses, and tried to identify areas where methodological development is needed. In doing this I have tried not to let my own biases influence the process of collecting and analysing data, and in forming conclusions. Inevitably though, I am not completely divorced from the research, and cannot ignore the fact that my efforts are likely to shape the world in which I exist. Indeed, texts on research methods in the social sciences acknowledge the undeniable role of the 'self' in scientific inquiry, noting that this is not necessarily a problem, providing there is some exploration of the researcher's involvement in the research (Hammersley and Atkinson, 2007; Nightingale and Cromby, 1999; Shacklock and Smyth, 1998). In the interests of transparency and personal reflexivity I have endeavoured to describe my role in this research throughout this thesis, and to document the methods used as clearly as possible.

Much of the literature on research interviewing stresses that the interviewer should be objective and impartial, to avoid influencing the interviewees' responses. As explained in Chapter 5, I was conscious not to let my status as an academic researcher with expertise in systematic reviewing in health promotion bias the comments made by the interviewees in this study. Although I have built up expertise over the last decade or so in this area (as mentioned in the Introduction to this thesis), modesty prevailing, I do not consider my status to necessarily have influenced their remarks. As far as I was aware the interviewees were not familiar with my work prior to the interview, and based on my experience of conducting the interviews and subsequent analysis of the data, I did not get the impression that they had over- or under-emphasised

particular issues on my account. Nonetheless, it is possible that they may have formed an impression about me prior to, or during the interview, and that this may have shaped what they told me.

In terms of personal reflexivity it is important to consider what impact this research may have had upon myself, my values and my work as a systematic reviewer. Having reported, digested and discussed the findings I consider that the research has considerably broadened my view of evidence, systematic reviewing, and decision making. It has provided me with a valuable opportunity to think beyond the confines of the day to day work of systematic reviewing, to consider broader issues concerning issues such as training, learning, research funding and capacity. Greater awareness of these issues has been invaluable given my increasing responsibility as a research manager within my department. Furthermore, having proposed a number of recommendations for the conduct of systematic reviews I am aware of the potential paradox of not living up to the expectations that my research has set. Whilst I hope the recommendations made are not unrealistic, I am nonetheless aware of the pressures involved in meeting the standards for, and increasing expectations of, systematic reviews.

Writing this thesis has also re-acquainted me of some of the philosophical and epistemological issues about the nature of knowledge generation and inquiry (e.g. positivism, realism) and paradigm debates in evaluation (e.g. quantitative and qualitative methods). During the day to day task of producing research these issues are often overlooked, but are nonetheless important in order to explain and contextualise our work.

8.2 The methodological framework

When completing any research project it is important to evaluate the methodological approach used and consider what could have been done differently. As explained in Chapter 2, this research used a multi method approach combining quantitative and qualitative data collection. The intention was to assess consensus by examining what methods systematic reviews have actually used (through the methodological mapping in Stage 1) and to ask a sample of systematic reviewers to discuss in greater detail the methods they have used, why they have used them, and their wider views on evidence synthesis and decision making (through the semi-structured interviews in Stage 2). Some of the research objectives were unique to a particular stage, whilst others were applicable to both (see Table 3 in Chapter 1). Therefore, to some extent the two research stages were independent of each other. However, it was intended that some of the key trends and themes arising from Stage 1 could be followed-up by the interviews in Stage 2, and that the stages should therefore be sequential. As mentioned in Chapter 5, rather

than being completely sequential there was some overlap between the two stages, due to slight delay in completing Stage 1 and Stage 2 being brought forward so that I could take advantage of the public health focus of the 2005 Cochrane Colloquium to collect data. There were a small number of issues that came to light when I wrote the discussion chapter of this thesis that, with hindsight, it would have been interesting to explore in the interviews. It would therefore have been advantageous to have completed collecting, analysing and digesting the data in Stage 1 before conducting the interviews in Stage 2. That said, the interviews themselves generated a wealth of data and there was certainly no shortage of issues and themes to discuss. In summary, on reflection I consider that the methodological framework used in this study was generally appropriate to enable the overall aim of the research, and the research objectives to be met.

8.3 The scope and contribution of the research

My original ideas for this research were focused on the issue of quality assessment in systematic reviews of the effectiveness of health promotion. Whilst this has remained a central concern of the research, my ideas have broadened to incorporate wider issues around the strengths and weaknesses of reviews and evidence-based health in general, and skills and learning for systematic reviewing. This diversification was prompted not only by the review of the literature, and the useful ideas from the experts interviewed as part of the agenda-setting exercise, but also (and inevitably) by my own changing experiences and ideas as a researcher actively engaged in this field. Such diversification of research ideas is common in all forms of research and illustrates natural curiosity on the part of the researcher to explore unanticipated and emergent issues.

This study has therefore covered a variety of issues, but perhaps not all of them to the level of detail that they potentially could be investigated. For example, the issue of external validity in health promotion has received comparatively little attention compared to internal validity, and it could be the subject of a thesis of its own. A great many questions have been posed, and in attempting to answer these a number of other important questions have emerged. This study could therefore be viewed as a starting point for further investigation of these and other issues. Indeed, perhaps one of the biggest contributions this investigation has made is the generation of a number of research recommendations, as summarised in Chapter 9.

8.4 The subjects of this research

Another one of the strengths of this research is the voice it has given to those who conduct systematic reviews. While much has been written about systematic reviews and evidence-based health there are very few published accounts, particularly in health promotion, of the

experiences and reflections of systematic reviewers (Nind, 2006; Oakley, 2003; Wallace *et al*, 2006). Without this study their views would remain anecdotal. However, it should be noted that the interviewees were not all exclusively systematic reviewers in health promotion. Although all had conducted at least one systematic review of a health promotion topic, they varied in terms of their expertise in this area. Whilst some had specialised in health promotion and public health for a number of years, in the main the interviewees did not consider themselves experts in this area and a couple of them remarked that they did not think their comments were necessarily representative of health promotion. It is perhaps not surprising that there were fewer reviewers with extensive experience of health promotion, given that health promotion and systematic reviewing are both relatively specialised fields. However, one of the advantages of sampling interviewees with experience of both health promotion and health care is that they were able to give a balanced view, and provide their perspectives on issues specific to health promotion, but also issues relating to health more generally. The write up of the results of the interviews (Chapter 6) has endeavoured to reflect the differences in expertise, noting the experience of the interviewees where appropriate.

It should also be acknowledged that this study is limited in what it can conclude about the views of other stakeholders, such as policy makers, health care consumers and practitioners (with the exception of practitioners who were also academics). Further research in this area could build on the current study by investigating their views and contrast with those of systematic reviewers.

8.5 Data collection

The strength of the methodological mapping exercise was that it was a systematic and transparent process in which a set of standardised data were extracted from each review. The potential shortcoming was, unlike standard procedure in systematic reviews, that this was performed by myself without independent verification from a co-reviewer. However, involvement of a co-reviewer was not possible in what was an independent study that received limited funding. Despite attempts to ensure accuracy and fairness in the extraction and interpretation of data it must be acknowledged that some subjective judgement is inevitable. Indeed, one of the key findings of this study is that subjectivity cannot be avoided during this kind of systematic investigation.

The mapping itself was performed on what could be considered to be a relatively small sample of reviews. Deciding on an optimum number of reviews to include is an arbitrary process. At the start of the study I had a notional limit of 50 reviews to include, but saturation was

considered to have been achieved by around 30 reviews. Some of the other methodological mapping studies cited in this thesis (e.g. Moja *et al*, 2005) included several hundred reviews. However, these studies were larger-scale funded pieces of research conducted by teams of people, none of them were focussed solely on health promotion, and none extracted qualitative data from the reviews. Whilst the sample in this study is perhaps too low to perform meaningful statistical testing it was able to identify trends, and afforded a detailed textual analysis of the views of the authors of the systematic reviews. A further, larger, study would be useful to confirm the trends identified here.

8.6 Chapter summary

This chapter has provided a brief critique of the strengths and weaknesses of this study. In doing this I have given a reflexive account of the process of conducting the research, discussing issues such as the methods used for data collection, and the contribution this research makes to the field. The next chapter expands on the latter with conclusions and a number of specific recommendations for researchers, research funders and the wider academic community.

Chapter 9 – Conclusions and recommendations

Few published investigations have covered in any great detail the areas of consensus and dissension around the production of systematic reviews of health promotion. This study has aimed to fill this gap.

9.1 Cross-cutting themes

Three key themes emerged from the results of this study: complexity of the evidence base, the subjective judgement needed to appraise evidence, and the pressure on time and funds. These were recurring and inter-linked themes that permeate many of the issues discussed.

- In terms of complexity, the issues discussed are not unique to health promotion, but some are certainly more acute in this area. Health promotion is an activity often characterised by multi-component and multi-disciplinary interventions designed to effect change at the level of policy, community, and individual. It requires a multiplicity of evaluation designs and sophisticated methods of analysis, placing additional demands on systematic reviewers in terms of expertise, time and resources.
- Despite explicit procedures and criteria, systematic reviewing often requires an element of subjective judgement and discretion, particularly in the appraisal of quality. This can be disconcerting particularly for novice reviewers who lack confidence to deviate from the *modus operandi*.
- The resources needed to produce a credible review have increased significantly over the years, and pressure on time and funding can generate a number of challenges. It can prevent reviewers from doing a review to an acceptable standard, it can limit the extent to which reviewers can adequately deal with the numerous shortcomings in the reporting of primary studies; it can be a barrier for busy practitioners, policy makers and other representatives of stakeholders to participate; and it may potentially represent a significant opportunity cost for academics under pressure to meet research targets.

This study has teased out some potentially promising solutions to these challenges which have been attempted by the systematic reviewers interviewed. Some empirical methodological research has been published in the field, but more remains to be done, particularly to clarify conflicting findings. Further research and development is needed to formally address these challenges, and specific recommendations have been made throughout this chapter. Where possible these recommendations will be made to relevant research funders, particularly the MRC Methodology Programme.

9.2 Key conclusions

The results of this study show that those who produce systematic reviews consider them to have many strengths, but they also acknowledge their weaknesses. Some of the long-standing debates about hierarchies of evidence continue, but judging from the comments made and from the literature analysed, much effort has gone into achieving progress. Initiatives to raise standards of reporting in both primary and secondary research such as the TREND and QUOROM (Quality of Reporting of Meta-Analysis) statements will hopefully make evidence more accessible to all. Systematic reviews, particularly those conducted as part of a policy initiative such as health technology assessment, appear to be informing policy and practice. Infrastructures exist for placing the gaps in evidence base that systematic reviews identify on the research agenda, and a great deal of investment has been made to commission high quality primary and secondary research.

This study has found that, in the main, the conduct of systematic reviews of health promotion accords with recommendations from guidelines and texts on evidence-based health. Quality assessment is routinely performed in systematic reviews of health promotion, albeit to varying degrees, and it is considered one of its defining characteristics. Quality is considered at all key stages of a review, most commonly during the synthesis of results, during the process of screening studies for inclusion, and in the discussion of the findings. The vast majority of systematic reviews incorporate their judgement of the evidence into account in the analysis of effectiveness, as recommended. This is predominantly done qualitatively, as opposed to using 'empirical' (quantitative) methods such as sensitivity analysis. However, there was little discussion of, or justification for the methods used, and this perhaps reflects the fact that there does not appear to be clear guidance on which method(s) should be used. Alarmingly, half of the reviews in this study were judged to be inconsistent or ambiguous in their approach, typically critiquing studies without reporting any formal assessment of quality, or precise details of the criteria employed.

The criteria used to assess the quality of evidence in health promotion systematic reviews are diverse. There was some degree of consensus from published systematic reviews that RCTs and experimental evaluations are the favoured design. RCTs and experimental designs were permitted by the majority of reviews, although not exclusively so, and not necessarily without consideration of their weaknesses. Observational designs were included where appropriate and were considered to have merits in particular circumstances. The criteria employed by systematic reviews of health promotion include many of the attributes commonly used to assess the quality of health care interventions, including blinding of study group assignment, often considered to

be impossible in health promotion. Criteria that could be considered particularly relevant to health promotion, such the adequacy of the length of follow-up or the potential for contamination from competing interventions, are also used but to a lesser extent. Justifications for choice of criteria vary and, despite what appears to be a general awareness of empirically-established key threats to internal validity, supporting literature is not always cited.

The prevailing view amongst those interviewed was that there was probably no or little consensus about quality assessment criteria in health promotion. Furthermore, to expect consensus in such a multi-disciplinary area was considered unlikely. Methodological pluralism is evident, although steps are being taken to foster collaboration and a shared ways of working in order to reduce duplication of resources, and encourage consistency in policy making. Whilst it may not be realistic to expect complete standardisation of methods, there are encouraging developments in empirical methodological research which may lead to greater clarity around some of the methodological uncertainties.

It is encouraging that the vast majority of systematic reviews in this investigation assessed the external validity of included studies, although few report conducting a formal appraisal. Most commonly reviews addressed the generalisability and replicability of the evidence, such as the delivery of the intervention and the characteristics of the study population. However, more research is needed to devise methods for considering cost-effectiveness of health promotion, and markers of the quality of the intervention to facilitate the production of guidance by policy making organisations such as NICE. This study has made inroads into this relatively neglected area but further, more detailed, investigation is required.

It would appear that systematic reviewing is an academic activity performed mostly by researchers (although sometimes they may be academic practitioners), with little collaboration with other stakeholders, particularly lay people. This is at odds with current health policy which strives to involve practitioners and consumers of health services in all aspects of research. The reviews included in this investigation tended to be conducted by people with expertise in the topic under review. However, in a small number of cases reviews were the product of multi-disciplinary teams of people and, despite some drawbacks, this is considered by systematic reviewers to be a useful way forward to compensate for the obvious disadvantages of single disciplinary reviewing.

People learn to do reviews through a combination of approaches including training, reference to written resources, hands-on experience and support from mentors. Practice-based learning in particular is considered to be very helpful. There has been a proliferation of texts on the

methodology of systematic reviewing and training courses, and in the main these are considered adequate, although greater attention to complex interventions is considered necessary. The philosophy and practice of systematic reviewing is now an integral part of many higher degrees in health, suggesting that future generations of health professionals will possess some competence in critical appraisal and other skills necessary for systematic reviewing. The fact that some novices find particular aspects of systematic reviewing a challenge to master casts doubt upon the ability of some, notably those without an academic background, to acquire such these skills, and to be taken seriously as systematic reviewers. This may become less of a problem if a team approach to systematic reviewing is to become the norm, and there are signs that training programmes are catering for those in this position.

9.3 The Future

The overall conclusion of this thesis is that whilst areas of dissension remain in the production of systematic reviews, there is evidence of progress in the pursuit of consensus and methodological advancement to overcome challenges. It is hoped that there the current infrastructure will be maintained and expanded to support the production of adequately resourced systematic reviews of health promotion, addressing policy-relevant questions, based on sound methodology, cognisant with up-to-date methodological guidelines, produced with input from all relevant stakeholders, and effectively disseminated. This will contribute to the goal of effective health promotion, and ultimately to health gains.

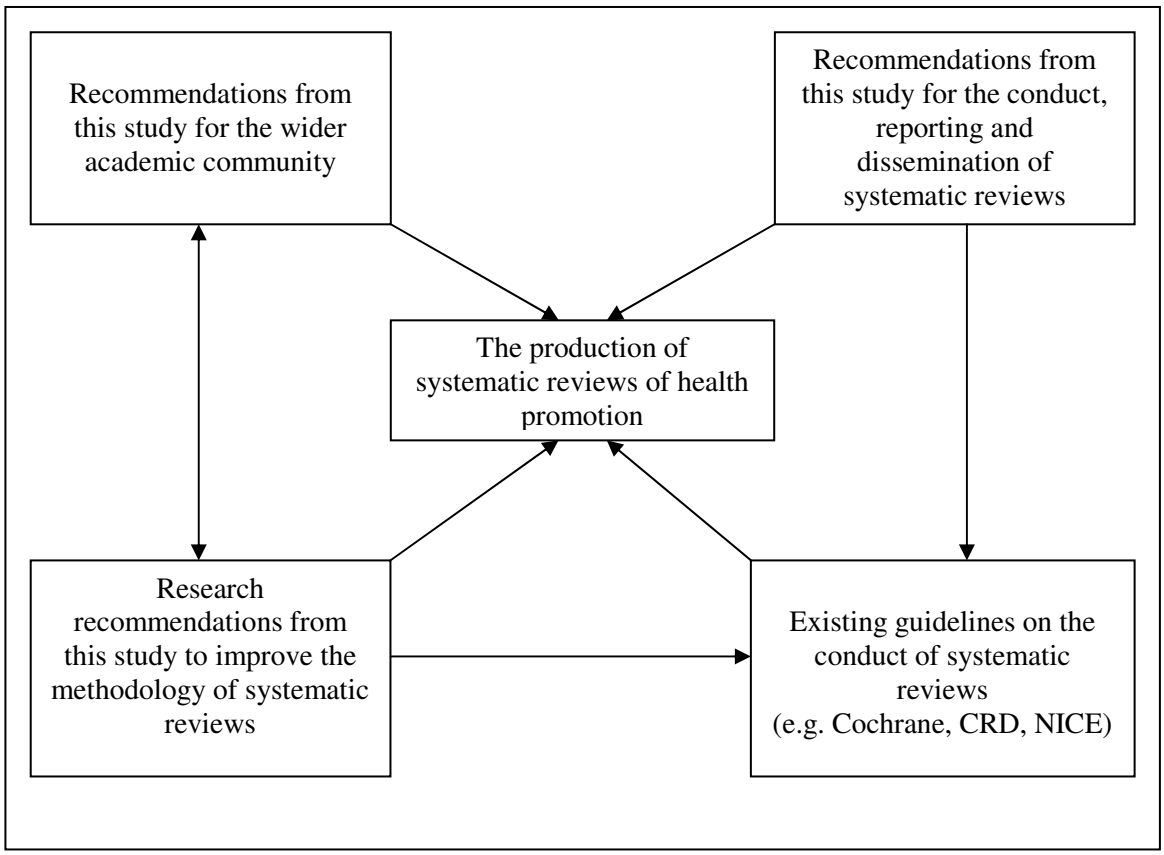
9.4 Summary of recommendations

The recommendations made in Chapter 7 are summarised below according to their relevance to the conduct of systematic reviews, to further research, to the wider academic community. Figure 14 shows the inter-relationships between the three different sets of recommendations and with existing guidelines on the production of systematic reviews (e.g. The Cochrane Handbook; the CRD guidance for undertaking systematic reviews; NICE's guide to the methods for developing public health guidance). The guidelines make recommendations for the production of systematic reviews, many of which are reinforced and advanced by the findings of this study. This study also makes recommendations for further research to improve the methodology of systematic reviews, which should be incorporated into future updates of the guidelines, and in turn be taken up by systematic reviewers.

Recommendations are also made by this study for the wider academic community, which are intended to influence the policy context within which systematic reviews are commissioned, produced and disseminated. It is my intention to actively disseminate these recommendations

through articles in academic journals, conference presentations, and informally through my professional networks (e.g. via the Cochrane Collaboration, CRD, NICE, the EPPI-Centre). I have submitted an abstract to the 2009 Cochrane Colloquium to present the results of this research (outcome pending) (Shepherd, 2009)

Figure 14 – The inter-relationships between recommendations from this study, and guidelines on the production of systematic reviews of health promotion



9.4.1 Recommendations for the conduct of systematic reviews

This section contains recommendations arising from this thesis for the conduct of systematic reviews.

1. It is important for systematic reviews to be realistically planned to ensure that adequate time is available for a thorough and fair assessment of quality, particularly to allow for contacting study authors for clarification in cases of poorly reporting.
2. Systematic reviews should explicitly justify their choice of quality assessment criteria with reference to up to date empirical evidence demonstrating how the criteria account for bias.

3. The process of quality assessment, including the criteria used should be explicitly reported, so that the basis of their judgement is transparent allowing users of reviews to determine whether or not they are fair and the results of the review are therefore credible.
4. Systematic reviews should use a combination of approaches to incorporate quality judgement into the synthesis of evidence, and provide a justification for the methods they have chosen.
5. Systematic reviews of health promotion should assess external validity by using a framework designed to gauge the potential for generalisability and replicability. It is also recommended that systematic reviews routinely assess the quality of the interventions included. In particular they should assess the appropriateness of theories cited by studies, to explain the mechanism through which interventions are purported to work. Guidance on the most appropriate markers of ethical soundness is needed to assist reviewers in handling studies which may be ethically dubious. Assessment of intervention quality should be informed by further research on these and other markers of quality (See recommendation 15 below)
6. Given the increased focus on demonstrating cost-effectiveness in policy making it is recommended that systematic reviews of health promotion routinely assess the costs and resources of health promotion, subject to adequate capacity in terms of expertise in health economics (See recommendation 20 below).
7. Systematic reviews should be conducted by multi-disciplinary teams, where possible, involving all relevant stakeholders. Teams should be carefully monitored and evaluated to ensure efficiency and fairness in the judgement of the quality of evidence.
8. It is recommended that people wishing to learn to conduct systematic reviews be given the opportunity to have hands on experience, in conjunction with other learning strategies as appropriate.
9. It is essential that guidelines on systematic reviewing are updated regularly to take into account new developments in methodological research (such as empirical research into bias – see Section 9.4.2 below), to provide systematic reviewers with clarity around issues such as quality assessment criteria.

9.4.2 Recommendations for further research

Below are specific recommendations for methodological research arising from this thesis. The purpose of the research is to improve the conduct of systematic reviews, and should feed into regularly updated guidelines on systematic reviewing.

10. It is recommended that the impact of systematic reviews of health promotion be evaluated on an on-going basis to ensure that the needs of all stakeholders are being met (e.g. that they

are making effective contributions to decision making). This is particularly important in terms of public accountability, given government commitments to increase funding for evidence synthesis in this area.

11. Further empirical methodological research, including tertiary reviews, should be commissioned investigating factors related to bias, to try and reconcile divergent findings of existing studies. This will reduce the uncertainty amongst systematic reviewers about the relative importance of different biases to consider when assessing quality.
12. Methodological research should be conducted with a view to reconciling debates about the appropriateness of including observational studies in the absence of experimental designs**.** This might take the form of further qualitative research.
13. Investigation into effective ways to foster collaboration and achieve consensus in terms of methods of systematic reviewing would be beneficial. This would reduce duplication of effort (to ensure efficient use of public funds) and to lessen the risk of conflicting findings from systematic reviews and the negative knock-on implications for policy and practice.
14. Further investigation should assess the advantages and disadvantages of different approaches to incorporating quality assessment judgement into the analysis of results of a systematic review (e.g. weighting, sensitivity analysis). Where possible strategies should be empirically tested to examine the impact on the overall results of systematic reviews.
15. Further research is needed to identify other markers of intervention quality in health promotion which could be employed by systematic reviews. Consensus development methods involving a range of stakeholders might be a useful way to prioritise quality markers.
16. It is recommended that research is conducted to assess the barriers and facilitators to involving all stakeholders in the production of systematic reviewers (e.g. practitioners, policy makers, the public), to inform evaluation of promising approaches to seeking stakeholder involvement.
17. Further investigation is needed to identify potentially effective ways of addressing the challenges faced by those who train systematic reviewers (e.g. around the explanation of statistics and quality assessment). Potentially effective strategies could be implemented and evaluated to assess their impact.
18. The innovative strategies elicited by this study for overcoming the challenges presented by poor reporting of primary studies should be subjected to empirical evaluation to assess the benefits in terms of a potentially more thorough and rigorous systematic review, and to establish whether strategies are realistic in terms of the time and resources required.

9.4.3 Recommendations for the wider academic community

This final section contains recommendations arising from this thesis for the commissioning, conduct and reporting of primary and secondary research.

19. An over-arching recommendation is that primary evaluations should report as much detail as possible on all aspects of the intervention and its delivery to enable systematic reviews to make a thorough assessment of internal and external validity. The extended CONSORT statement and the TREND statement on reporting standards should be adhered to by academic journals in the area of health promotion and public health. The effectiveness of these consensus statements and other initiatives to improve reporting of evaluation should be monitored.
20. Primary evaluations of health promotion interventions should, where necessary, conduct a an economic evaluation to meet the need for evidence of cost-effectiveness by policy makers. If a primary economic evaluation is not possible the resources used and costs incurred in providing the intervention should be reported in detail to enable a secondary economic evaluation to be conducted. More investment is required to ensure there is adequate capacity, in terms of systematic reviewers with at least a basic understanding of health economics, to meet the demand for evidence of cost-effectiveness.
21. It is recommended that academic journals, including those with a high impact factors, consider accepting systematic reviews for publication in their editorial policy, to provide greater incentives for senior academics to consider conducting and publishing systematic reviews.
22. It is recommended that the investment made in recent years in building an infrastructure for the production and dissemination of systematic reviews of health promotion be maintained. This will ensure that systematic reviews are commissioned and produced in a timely fashion and to a high standard. Funding opportunities for systematic reviews should be evaluated to ensure they are accessible to all, and that they represent the diversity of health promotion interventions and topics.
23. It is recommended that commitment to training is extended in terms of provision of bursaries and scholarships to increase access to courses, particularly for those not in a position to take advantage of training opportunities due to factors such as geographical location, and availability of time and funds. This should be done in conjunction with initiatives to increase wider stakeholder participation in systematic reviews.

Appendices

Appendix 1 – Methods for the agenda-setting interviews

A. 1 Rationale for the exercise

Although the research objectives proposed in Chapter 1 of this thesis were derived from a thorough examination of the literature, it was considered important to augment them with views from a small sample of systematic reviewers with experience of health promotion. Their views were sought on key issues which had not been identified from the literature and which issues to prioritise. In this respect the exercise can be viewed as a form of 'peer validation' of my research ideas. Although essentially a precursor to the project, the agenda-setting interviews can be considered as research in its own right. The exercise had two aims:

1. To seek views from a small sample of systematic reviewers on the aim of this study and the research objectives.
2. To seek views on whether the draft data extraction instrument to be used in Stage 1 of the research covered all the key issues relevant to this investigation.

The findings were used to refine the instrument and, along with the literature reviewed in Chapter 1, to fine tune the agenda for the study.

A.2 Methods

Brief semi-structured interviews intended to last between 15 to 30 minutes were planned with between five and ten systematic reviewers. A convenience sample was considered acceptable for this purpose and therefore I approached six people I knew professionally in the area of evidence synthesis, all of whom consented to be interviewed. None of these people were interviewed in Stage 2 of the study.

A copy of the draft data extraction instrument was sent to the interviewees beforehand, and formed the basis of the discussion during the interview. The interviews were not tape recorded, but notes were taken during the interview and typed up immediately after. Where requested the notes were later sent to the interviewees to confirm their accuracy. Once all interviews had been completed the notes were analysed to identify key issues to be grouped together under themes.

A.3 Copy of the interview schedule

Notes on what to say to the interviewee:

- Aim of this discussion: to assist in the development of data extraction tool to be used in Stage 1 of the research.
- Output – notes from discussion (and others) will be collated and written up. They will help to set the agenda for the study and inform the further development of the data extraction instrument. Any of the data from this interview will be unattributed in the thesis. Do you consent to this?
- Outline of my PhD (verbal). Two main stages, possibly a third. Today focusing on first stage. Sample of systematic reviews at random from DoPHER, apply instrument to each one.
- Aim of the tool: to map the methodological characteristics of health promotion systematic reviews (in its own right), in order to identify further issues for second stage of fieldwork – in-depth interviews with systematic reviewers. The questions that the tool asks have been derived as key issues to investigate from the literature. However I before actually using it I would like to ask a small sample of people working in this area for their ideas, views etc on key issues to focus on.

Questions

- Any questions or comments on the data extraction instrument?
- What do you see as being the key issues to look at?
- Which do you think might be less important?
- Which issues do you think might be missing?
- External validity
- Any comments on the structure?

Appendix 2 – Stage 1 fieldwork - Data extraction instrument

Section A: General Details

A.1 Funder	A.1.1 Stated (Specify) A.1.2 Not stated
A.2 Was a protocol for the review published?	A.2.1 No / Not stated A.2.2 Yes (specify details) *Specify: where published; whether or not peer reviewed*
A.3 Where is the final review published?	A.3.1 Peer-reviewed journal A.3.2 Report *Specify whether peer-reviewed report if possible* A.3.3 Book A.3.4 Cochrane library A.3.5 Unpublished A.3.6 Dissertation / Thesis *Specify whether Masters level or PhD* A.3.7 Other (Specify)
A.4 Focus of the report *Tick as many as apply*	A.4.1 accidents A.4.2 alcohol A.4.3 asthma A.4.4 cardiovascular A.4.5 cancer A.4.6 skin cancer A.4.7 child neglect A.4.8 delinquency A.4.9 diabetes A.4.10 disability A.4.11 disease A.4.12 drugs A.4.13 eating disorder A.4.14 education system A.4.15 emotional abuse A.4.16 epilepsy

	A.4.17 health promotion A.4.18 healthy eating A.4.19 hygiene A.4.20 inequalities A.4.21 injury A.4.22 leisure A.4.23 medical care A.4.24 mental health A.4.25 obesity A.4.26 oral health A.4.27 parenting A.4.28 physical abuse A.4.29 physical activity A.4.30 pregnancy prevention A.4.31 problem behaviour A.4.32 sexual abuse A.4.33 sexual health A.4.34 solvents A.4.35 STD A.4.36 Suicide A.4.37 tobacco A.4.38 workplace A.4.39 Other (Specify)
A.5 Type of intervention *Tick as many as apply*	A.5.1 Not stated A.5.2 Unclear A.5.3 Advice/counselling A.5.4 Anger management A.5.5 Bio-feedback *i.e. feedback to an individual their biological measure(s) and/or behavioural/social risk status indicating potential health consequences e.g. the level of carbon monoxide in the blood relating to smoking* A.5.6 Brief therapy A.5.7 Casework A.5.8 Environmental modification (please specify) A.5.9 Family therapy (please specify)

	A.5.10 Increased access to resources (specify) A.5.11 Increased access to services (specify) A.5.12 Information/education A.5.13 Legislation/regulation A.5.14 Parent training intervention A.5.15 Professional training A.5.16 Physical activity A.5.17 Practical skill development (specify) A.5.18 Risk assessment *The establishment of a risk profile (not solely relying on medical screening) for a particular adverse outcome, which is not fedback on an individual basis* A.5.19 Screening *medical screening (eg breast screening, ultrasound)* A.5.20 Social support A.5.21 Other (please specify)
A.6 What was the aim of the review?	A.6.1 Stated (specify) A.6.2 Not stated A.6.3 Unclear
A.7 Authors' qualitative description of intervention	A.7.1 Stated (specify) A.7.2 Not stated
A.8 Did the review have an advisory group / steering group?	A.8.1 Yes (Specify) A.8.2 No / Not stated

Section B: Quality assessment: general details

B.1 Are the inclusion criteria (as regards study methodology) specified? *Specify what type of study design the authors included (e.g. RCTs; CCTs; observational studies) and/or the specific methodological features of studies that authors considered necessary for a study to be included (e.g. only include studies with less than 30% attrition).*	B.1.1 Yes (Specify) B.1.2 No/not stated
B.2 Are the inclusion criteria (as regards type of evaluation) specified?	B.2.1 Yes B.2.2 No/Not stated

B.3 If inclusion criteria (as regards type of evaluation) is specified what evaluation types are used? *Tick as many as apply*	B.3.1 Process evaluation B.3.2 Outcome evaluation B.3.3 Other (Specify) B.3.4 Not applicable
B.4 Is quality assessment performed in this review?	B.4.1 Yes B.4.2 No
B.5 If quality is assessed at which stage of the review was it performed? / which section of the article is it discussed? *Tick as many as apply*	B.5.1 Inclusion/exclusion criteria B.5.2 Discrete quality assessment stage B.5.3 Synthesis B.5.4 Discussion / Recommendations B.5.5 Quality not assessed at all B.5.6 Unclear
B.6 Further comments on the stage at which quality is considered	B.6.1 Specify
B.7 How are the results of the quality assessment presented? *Tick as many as apply*	B.7.1 Narrative format *Using words to summarise quality of the studies (e.g. "Most of the studies were of high quality, and took steps to minimise selection and other biases").* B.7.2 Tabular *Where results are presented in the form of a table (e.g. a list of all the studies and whether or not they met each of the criteria)* B.7.3 Scorings/ratings *Whereby a numerical score is given to each study representing its quality (e.g. 5 out of 10); or whereby studies are ranked according to their quality.* *(NB. Both could be presented in the form of a table)* B.7.4 Not applicable (quality assessment not stated) B.7.5 Other (Specify)
B.8 What form of synthesis is employed? *Tick as many as apply*	B.8.1 Narrative synthesis B.8.2 Meta analysis B.8.3 Other (Specify)
B.9 Are assessments of quality integrated into the synthesis?	B.9.1 Yes B.9.2 No
B.10 If assessments of quality are integrated into synthesis what method is used?	B.10.1 Sensitivity analysis

	The authors report how the results vary according to the best and worse quality studies B.10.2 Threshold analysis *The authors only report the results of studies meeting a given threshold of quality (e.g. scoring above 7 out of 10)* B.10.3 Narratively *The authors report the results of the review in the context of the quality of the study using words (e.g. in the results/discussion/conclusion sections)* B.10.4 Weighting *The authors give results of better quality studies more emphasis.* B.10.5 Other (Specify)
B.11 Further details on integration of quality into the synthesis (if appropriate)	B.11.1 Specify
B.12 If quality is not integrated into synthesis do the authors provide a reason for this?	B.12.1 Yes(specify) B.12.2 No B.12.3 Not applicable (quality is integrated into synthesis/quality assessment not reported)
B.13 Who was involved in assessing quality? *Tick as many as apply.* *Note whether it was a team effort, and whether the appraisers were topic specialists or systematic review experts (if it can be deduced).*	B.13.1 Researcher B.13.2 Lay person / consumer B.13.3 Student B.13.4 Practitioner B.13.5 Policy specialist B.13.6 Other B.13.7 Not stated B.13.8 Not applicable (quality assessment not reported)
B.14 Are judgements of study quality made by more than one person independently? *NB. This question still applies if quality is considered only as part of the inclusion criteria. You would want to know if more than one person screened studies for inclusion on methodological grounds.*	B.14.1 Yes (specify) *Specify if* *2* *3-4* *etc* B.14.2 No B.14.3 Unclear B.14.4 Not stated B.14.5 Not applicable (quality assessment not stated)

B.15 What is the background of the appraiser? *It is unlikely that this will be reported explicitly. A judgement may have to be made according to the stated designation of the authors. Where an inference is made please report this.*	B.15.1 Topic specialist B.15.2 Systematic reviewer B.15.3 Can't tell/ not stated B.15.4 Not applicable (quality assessment not reported)
B.16 What training/preparation were they given?	B.16.1 Stated (specify) B.16.2 Not stated B.16.3 Not applicable (quality assessment not reported)
B.17 Did the authors indicate any barriers to the process of quality assessment? *Tick as many as apply*	B.17.1 Not required by funding body B.17.2 Political constraints B.17.3 Lack of training/ experience B.17.4 Lack of reported detail on quality markers B.17.5 Other (specify) B.17.6 No barriers stated B.17.7 Not applicable (quality assessment not reported)
B.18 Did the authors indicate any facilitating factors to the process of quality assessment? *Tick as many as apply*	B.18.1 Critical appraisal training provided B.18.2 Further (unpublished) information provided by study authors B.18.3 Professional support (e.g. from Cochrane review group) B.18.4 Not required to by funding body/institution B.18.5 Other (specify) B.18.6 No facilitators stated B.18.7 Not applicable (quality assessment not reported)
B.19 Do the authors make recommendations for future evaluation methodology based on their quality assessment of the evidence?	B.19.1 Yes (specify) B.19.2 No B.19.3 Not applicable (quality assessment not reported)
B.20 Other details about the review / quality assessment process	B.20.1 Specify
B.21 If quality is not considered at all do the authors state why?	B.21.1 Yes (specify) B.21.2 No / Not applicable

Section C: Quality assessment criteria: internal validity

NB. There is no need to complete this section if answer to question B3 (At what stage of the review is quality assessed?) is B3.5 (Quality not assessed at all).

C.1 What are the criteria for assessing internal validity (DESIGN CRITERIA) *Tick as many as apply*	C.1.1 Post test only, 1 group C.1.2 Post test only, >1 group C.1.3 Cohort study C.1.4 Case control study C.1.5 Randomised controlled trial C.1.6 Controlled trial (non-random) / quasi-experimental C.1.7 One group pre and post C.1.8 Case series C.1.9 Interrupted time series C.1.10 Case study C.1.11 Other (specify) C.1.12 None reported
C.2 What are the criteria for assessing internal validity (CRITERIA RELEVANT TO CONTROLLED TRIALS ONLY) *Tick as many as apply*	C.2.1 Other (specify) C.2.2 Method of allocation *Applies to both randomised and non-randomised controlled trials. For RCTs it refers to the validity of the randomisation method (i.e. is it really random?). For CCTs it refers to attempts made by the authors to minimise selection bias by techniques such as matching study groups.* C.2.3 Equivalent baseline study groups / adjustment for inequivalence C.2.4 Reports number of people in each study group C.2.5 Blinding *make a note as to whether single, double or triple blinded* C.2.6 Concealment of allocation process C.2.7 None reported
C.3 What are the criteria for assessing internal validity? (CRITERIA RELEVANT TO CONTROLLED TRIALS AND OTHER DESIGNS) *Tick as many as apply*	C.3.1 Statistical power calculation/ sample size C.3.2 Validity and reliability of data collection instruments/methods *Specify: validated instrument; assurance of confidentiality/anonymity; trained*

	interviewers; C.3.3 Validity and reliability of data analysis methods *Including: whether the unit of analysis matches the unit of assignment; whether an 'intention to intervene' or an 'intervention received' analysis was performed; whether cluster trials are analysed correctly.* C.3.4 Outcome measures / All outcomes reported on C.3.5 Clearly defined aims C.3.6 Attrition/ Loss to follow-up discussed C.3.7 Pre-intervention data provided C.3.8 Post-intervention data provided C.3.9 Length of follow-up C.3.10 Informed consent *NB. not necessarily related to internal validity, but is a marker of quality.* C.3.11 Findings support conclusions *i.e. do the conclusions of the included studies reflect the results presented?* C.3.12 Hawthorne effect / testing effect *Note that "Hawthorne" is not the name of a researcher, but of the factory where the effect was first observed and described: the Hawthorne works of the Western Electric Company in Chicago.* *One definition of the Hawthorne effect is: An experimental effect in the direction expected but not for the reason expected; i.e. a significant positive effect that turns out to have no causal basis in the theoretical motivation for the intervention, but is apparently due to the effect on the participants of knowing themselves to be studied in connection with the outcomes measured.* C.3.13 Contamination / co-intervention *Where the effects associated with the experimental intervention may have occurred because of another intervention that took place (e.g. mass media campaigns).* *Or, where the control/comparison group receive the experimental intervention* C.3.14 None reported C.3.15 Other (specify)

C.4 What are the criteria for assessing internal validity? (OTHER)	C.4.1 Specify
C.5 Authors' qualitative description of criteria	C.5.1 Specify
C.6 What justification is provided for the criteria used?	C.6.1 Criteria supported by empirical evidence on protection against bias (Specify) *Record references to any empirical studies* C.6.2 Criteria have been used in other systematic reviews (Specify) *Record references to any reviews* C.6.3 Criteria are recommended by systematic review methodology guidelines (Specify) *Record any references to guidelines* C.6.4 No justification given C.6.5 Other (Specify)
C.7 Other details about the internal validity criteria	C.7.1 Specify

Section D: Quality assessment instrument

NB. There is no need to complete this section if answer to question B3 (At what stage of the review is quality assessed?) is B3.5 (Quality not assessed at all).

D.1 Number of items *An item is a specific question/criterion (e.g. 'Were participants aware of which group they had been allocated to?'). If possible state the number of items that relate to internal validity and external validity, respectively.*	D.1.1 Stated (Specify) D.1.2 Not stated
D.2 What kind of instrument was used to assess quality?	D.2.1 Scale D.2.2 Other (Specify) D.2.3 Not stated D.2.4 Checklist D.2.5 Not relevant (quality not assessed / quality specified in inclusion criteria)
D.3 Number of components *A component is a group of items (e.g. section comprising questions/criteria relating to blinding). If possible state the number of components that relate to internal validity and external validity, respectively.*	D.3.1 Stated (Specify) D.3.2 Not stated
D.4 Did the authors describe how this had been developed/piloted/validated?	D.4.1 Yes (specify)

	D.4.2 No
D.5 Has the instrument been used in a previously cited review/study?	D.5.1 Yes (specify) D.5.2 No / not stated
D.6 What is the name of the instrument?	D.6.1 Stated (specify) D.6.2 Not stated
D.7 Other details about the quality assessment instrument	D.7.1 Specify

Section E: Quality assessment criteria: external validity

E.1 For what purpose does the review address external validity? *Tick as many as apply*	E.1.1 To explain results (e.g. to identify predictors of outcome) *e.g. through process evaluation to identify what factors contributed to success/failure; and/or through statistical procedures (e.g. multivariate regression analyses) to identify significant interactions between independent variables (e.g. different aspects of the intervention) and dependent variables (outcomes).* E.1.2 Other (Specify) E.1.3 To assess quality *i.e. to provide some evaluative judgement about the intervention and how it was devised, planned, and delivered.* E.1.4 To provide context within which to interpret outcomes *i.e. the reviewer extracts and presents data on the intervention and the study group primarily to give the reader background information.* E.1.5 To facilitate generalisability/replicability *i.e. to enable users of the review to gauge to what extent the results may be applicable to their location and/or to be able to replicate the intervention locally.*
E.2 What aspects of REPLICABILITY are assessed/extracted? *Tick as many as apply* *Replicability defined as the ability to reproduce the same intervention in a different location*	E.2.1 Intervention content (e.g. information provision, skills training, health care, access to resources, legislation, policy) E.2.2 Infrastructure (e.g. funding, costs, resources, organisation, planning) E.2.3 Other (specify)

	E.2.4 Replicability not assessed/extracted E.2.5 Intervention delivery (e.g. provider, setting, media, format, duration)
E.3 What aspects of GENERALISABILITY are assessed/extracted? *Tick as many as apply* *Generalisability defined as aspects of the study reported to enable others to judge whether the results may be applicable to their local populations/locations*	E.3.1 Outcome measures E.3.2 Details of population (age, sex, sexuality, ethnicity, culture, socio-economic status, location) E.3.3 Other (specify) E.3.4 Generalisability not assessed/extracted
E.4 What other aspects of the INTERVENTION are assessed/extracted?	E.4.1 Theoretical basis E.4.2 Based on a needs assessment E.4.3 Designed with input from target population E.4.4 Piloted E.4.5 Involvement of key stakeholders E.4.6 Other (specify)
E.5 Is any justification provided for issues assessed/extracted?	E.5.1 Yes (specify) E.5.2 No
E.6 Authors' qualitative description	E.6.1 Specify

Section F: JS comments on this review

F.1 Comment	F.1.1 Specify

Appendix 3 - Bibliography of the 30 reviews included in the methodological mapping (Stage 1)

N=21 systematic reviews included from the random sample of 50 reviews (taken in November 2003)

Aldana, S. G. & Pronk, N. P. 2001, "Health promotion programs, modifiable health risks, and employee absenteeism", *Journal of Occupational and Environmental Medicine*, vol. 43, no. 1, pp. 36-46.

Booth RE & Watters JK 1994, "How effective are risk-reduction interventions targeting injecting drug users?", *AIDS*, vol. 8, pp. 1515-1524.

Burke, L. E., Dunbar-Jacob, J. M., & Hill, M. N. 1997, "Compliance with cardiovascular disease prevention strategies: A review of the research", *Annals of Behavioral Medicine*, vol. 19, no. 3, pp. 239-263.

Campbell, M. 2000b, "A systematic review of the effectiveness of environmental awareness interventions", *Canadian Journal of Public Health*, vol. 91, no. 2, pp. 137-143.

Ciliska, D., Miles, E., OBrien, M. A., Turl, C., Tomasik, H. H., Donovan, U., & Beyers, J. 2000, "Effectiveness of community-based interventions to increase fruit and vegetable consumption", *Journal of Nutrition Education*, vol. 32, no. 6, pp. 341-352.

Cross, J. E., Saunders, C. M., & Bartelli, D. 1998, "The Effectiveness of Educational and Needle Exchange Programs: A Meta-Analysis of HIV Prevention Strategies for Injecting Drug Users", *Quality and Quantity*, vol. 32, pp. 165-180.

Dishman RK & Buckworth J 1996, "Increasing physical activity: a quantitative synthesis", *Med Sci Sports Exerc*, vol. 28, no. 6, pp. 706-719.

Dunn, A. L., Andersen, R. E., & Jakicic, J. M. 1998, "Lifestyle physical activity interventions. History, short- and long-term effects, and recommendations", *American Journal of Preventive Medicine*, vol. 15, no. 4, pp. 398-412.

Elders, L. A., van der Beek, A. J., & Burdorf, A. 2000, "Return to work after sickness absence due to back disorders--a systematic review on intervention strategies", *International Archives of Occupational and Environmental Health*, vol. 73, no. 5, pp. 339-348.

Fletcher A & Rake C 1998, *Effectivness of interventions to promote healthy eating in elderly people living in the community: a review*, Health Education Authority, Effectiveness Review nr. 8.

Foxcroft DR, Lister-Sharp D, & Lowe G 1997, "Alcohol misuse prevention for young people: a systematic review reveals methodological concerns and lack of reliable evidence of effectiveness", *Addiction*, vol. 92, no. 5, pp. 531-537.

Glanz K, Sorensen G, & Farmer A 1996, "The Heatlh Impact of Worksite Nutrition and Cholesterol Intervention Programs", *American Journal of Health Promotion*, vol. 10, no. 6, pp. 453-470.

Holtgrave, D. R., Qualls, N. L., Curran, J. W., Valdiserri, R. O., Guinan, M. E., & Parra, W. C. 1995, "An overview of the effectiveness and efficiency of HIV prevention programs.", *Public Health Reports.*, vol. 110, no. 2, pp. 134-146.

Hursti UK & Sjoden P 1997, "Changing food habits in children and adolescents: Experiences from intervention studies", *Scandinavian Journal of Nutrition*, vol. 41, pp. 102-110.

Rotheram-Borus, M. J., Cantwell, S., & Newman, P. A. 2000, "HIV prevention programs with heterosexuals", *AIDS*, vol. 14, no. 2 Supplement, p. S59-S67.

Silagy C, Mant D, Fowler G, & Lancaster T 1997, "Nicotine Replacement Therapy for Smoking Cessation," in *Tobacco Addiction Module of The Cochrane Database of Systematic Reviews The Cochrane Library [database on disk and CDROM]. The Cochrane Collaboration; Issue 4 - 1 Sep 1997*, Lancaster T, Silagy C, & Fullerton D (eds.), eds., Update Software, Oxford.

Snell JL & Buck EL 1996, "Increasing Cancer Screening: A Meta-Analysis", *Preventive Medicine*, vol. 25, pp. 702-707.

Stout J & Rivara F 1989, "Schools sex education: does it work?", *Pediatrics*, vol. 83, no. 3, pp. 375-379.

Thompson EL 1978, "Smoking Education Programs 1960-76", *American Journal of Public Health*, vol. 68, no. 3, pp. 250-257.

Turner JA, Schroth WS, & Fordyce WE 1996, "Educational and behavioural interventions for back pain in primary care", *Spine*, vol. 21, pp. 2851-2859.

Wolitski RJ, MacGowan RJ, Higgins DL, & Jorgensen CM 1997, "The effects of HIV counseling and testing on risk-related practices and help-seeking behavior", *AIDS Education and Prevention*, vol. 9, no. supplement B, pp. 52-67.

N=9 systematic reviews included from the random sample of 10 reviews (taken in October 2007)

Bambra, C., Whitehead, M., & Hamilton, V. 2005, "Does 'welfare-to-work' work? A systematic review of the effectiveness of the UK's welfare-to-work programmes for people with a disability or chronic illness.", *Social Science & Medicine*, vol. 60, no. 9, pp. 1905-1918.

Dyson, L., McCormick, F., & Renfrew, M. J. 2005, "Interventions for promoting the initiation of breastfeeding.", *Cochrane Database of Systematic Reviews* no. 2, p. CD001688.

Hillsdon, M., Foster, C., & Thorogood, M. 2005, "Interventions for promoting physical activity.", *Cochrane Database of Systematic Reviews* no. 1, p. CD003180.

Huibers, M. J., Beurskens, A. J., Bleijenberg, G., & van Schayck, C. P. 2003, "The effectiveness of psychosocial interventions delivered by general practitioners.", *Cochrane Database of Systematic Reviews* no. 2, p. CD003494.

Kirby, D., Laris, B. A., & Rolleri, L. 2006, *The Impact of Sex and HIV Education Programs in Schools and Communities on Sexual Behaviors among Young Adults*, Family Health International, YouthNet program, North Carolina.

Powell, C., Wedner, S., & Richardson, S. 2005, "Screening for correctable visual acuity deficits in school-age children and adolescents.", *Cochrane Database of Systematic Reviews* no. 1, p. CD005023.

Ogilvie, D., Egan, M., Hamilton, V., & Petticrew, M. 2004, "Promoting walking and cycling as an alternative to using cars: systematic review.[see comment].", *BMJ*, vol. 329, no. 7469, p. 763.

Shults, R. A., Elder, R. W., Sleet, D. A., Nichols, J. L., Alao, M. O., Carande-Kulis, V. G., Zaza, S., Sosin, D. M., Thompson, R. S., & Task Force on Community Preventive Services 2001, "Reviews of evidence regarding interventions to reduce alcohol-impaired driving.[erratum appears in Am J Prev Med 2002 Jul;23(1):72].", *American Journal of Preventive Medicine*, vol. 21, no. 4:Suppl, p. Suppl-88.

Tingle, L. R. 2003, "Short title A meta-evaluation of 11 school-based smoking prevention program", *Journal of School Health*, vol. 73, no. 2, pp. 64-67.

Appendix 4 – Key characteristics of the 30 systematic reviews included in the methodological mapping (Stage 1)

Table 4.1 - When the systematic reviews were published

When published	Number
2001- present	11
1995-2000	15
1991-1995	2
1986-1990	1
1981-1985	0
Pre 1980	1

Table 4.2 - Where the systematic reviews were published

Where published	Number
Peer-reviewed journal	23
Cochrane Database of Systematic Reviews	5
Report	2
Dissertation / thesis	0
Book	0

Table 4.3 - Topic areas covered by systematic reviews

Topic area	Number
Healthy eating	6
Sexually transmitted infections	6
Tobacco	6
Physical activity	5
Workplace health promotion	4
Cardiovascular health	3
Drugs	3
Health promotion in general	3
Sexual health	3
Accidents	3
Alcohol	3
Disability	3
Cancer	3

Pregnancy prevention	2
Injuries	2
Easting disorders	1
Hygiene	1
Problem behaviour	1
Asthma	1
Parenting	1
Health inequalities	1
Eye health	

NB. Systematic reviews could cover more than one topic, hence why total numbers exceed 30

Table 4.4 - Types of intervention

Type of intervention	Number of reviews
Information / Education	21
Advice / counselling	14
Practical skill development	12
Physical activity	8
Environmental modification	7
Increased access to resources	7
Increased access to services	7
Risk assessment	4
Screening	3
Legislation/regulation	3
Bio-feedback	3
Social support	2
Parent training	1
Professional training	1

NB. Systematic reviews could include more than one type of intervention, hence total numbers exceed 30

Table 4.5 - Systematic reviews that adapted existing published quality assessment criteria (n= sub-set of 5 reviews)

Systematic review	References to criteria adapted
Bambra *et al* (2005)	Croucher *et al* (2003) Edwards (2000) Egan *et al* (2004) Mays and Pope (2000)
Huibers *et al* (2003)	Jadad *et al* (1996) Verhagen *et al* (1998)
Foxcroft *et al* (1997)	Loevinsohn (1990) MacDonald *et al* (1992) Oakley *et al* (1995)
Ogilvie *et al* (2004; 2005)	Sustrans pilot project (2002)
Tingle *et al* (2003)	Boyd and Windsor (1993) Shufflebeam (2000) Sanders (1999)

Table 4.6 - Proportion of reviews citing different guidelines on systematic reviewing to support their choice of quality criteria (n=sub-set of 6 reviews)

Guidelines cited	Proportion of reviews N (%)
Cochrane Handbook (Higgins and Green, 2008)*	4 (66)
Joint Committee on Standards for Educational Evaluation (1994)	1 (17)
Cochrane Collaboration Back Review Group (van Tulder *et al*, 1997)	1 (17)
Agency for Health Care Policy and Research (1998)	1 (17)
Centre for Reviews and Dissemination (CRD) (2009)*	1 (17)
Cochrane Eyes and Vision Group Review Development Guidelines (2008)	1 (17)

NB. Reviews could cite more than one set of guidelines, hence why total numbers exceed 6

* The most recent versions of these reports are cited in the table, though the systematic reviews would have cited earlier versions.

Table 4.7 - Empirical texts cited by systematic reviews in relation to their quality assessment criteria (sub-set of 4 reviews)

Systematic review	Empirical study cited
Huibers *et al* (2003)	Schulz *et al* (1995)
	Chalmers *et al* (1983)
	Colditz *et al* (1989)
Tingle (2003)	Cohen (1988)
	Shufflebeam (2000)
	Biglan *et al* (1987)
Booth and Watters (1994)	Campbell and Stanley (1966)
	Cook and Campbell (1979)
Shults *et al* (2001)	Cooper and Hedges (1994)

Appendix 5 - Stage 2 Research – Final interview schedule

Instructions for interviewer:

This interview schedule is applicable to:

- People whose primary role involves conducting systematic reviews
- People whose primary role is not to conduct systematic reviews, but who have had some systematic review experience

Before starting the interview what you should say to the interviewee:

- I am going to ask you about your experiences on a range of issues relating to the production and use of systematic reviews, specifically in the area of health promotion and public health. One of the aspects of systematic reviewing that I am focusing on is the critical review, or quality assessment of the evidence.
- I will start off by asking you about your professional/research background, followed by some questions about your experience of conducting systematic reviews. We will then move on to discuss how you learned to do reviews, followed by some questions looking specifically at critical appraisal of research evidence.
- The interview will last for approximately 30 minutes, although it may be longer. All responses will be confidential, and that any quotes used will be anonymised. You can stop the interview at any time.
- The findings will be written up in my PhD thesis, but I also intend to publish the results in a peer reviewed journal, and/or at a conference. Please let me know if you would like to be kept informed of the results. You are welcome to see copy of the interview transcript if you wish.
- Are there any questions before I start?

Before starting the interview check:

- That the tape recorder is on, and the microphone appropriately positioned. That you have a back up ready to go if first malfunctions. Test the tape first.

A - Their professional background (warm up questions)

(Interviewer) "I am going to start off by asking you some general questions about your professional background"

A1. Could you tell me a bit about your current job/role
- *(prompt* e.g. Employer / institution)

A2. Could you summarise your current research interests?
(probe where appropriate)

B - Questions about systematic reviews

B1. Can you tell me a bit about what kind of research you have been involved in?
- *(probe e.g.* have you done primary research? What kind of data did you collect?)

B2. Could you tell me a bit about how you first became involved in systematic reviewing?
- (*prompt* e.g. what topics have you reviewed?)

B3. Which particular aspects of systematic reviewing have you had experience of?
- (*prompt e.g.* literature searching? Assessing quality?)
- *(probe where appropriate)* Can you tell me more about this?

B4. What aspects of it do you find the most interesting/satisfying, and which not?
- *(probe where appropriate)* Can you tell me more about this?

B5. What have been the biggest challenges you've faced in doing systematic reviews so far?
- *(prompt e.g.* Learning and applying the methodology)
- *(probe where appropriate)* Can you tell me more about this?

B6. How have you tried to deal with them?
- *(prompt e.g.* Seeking help from colleagues? Trial and error?)
- *(probe where appropriate)* Can you tell me more about this?

B7. With what degree of success would you say you dealt with them?
- *(probe where appropriate)* Can you tell me more about this?

B8. Is there anything else you want to tell me about your experience of systematic reviews?

B9.What do you see as being the strengths of systematic reviews?
- *(prompt)* Able to summarise large volumes of literature
- *(probe where appropriate)* Can you tell me more about this?

B10. What do you see as being the weaknesses of systematic reviews?
- *(prompt e.g.* Not user friendly, particularly for a non-research audience)
- *(probe where appropriate)* Can you tell me more about this?

C - Learning about systematic reviews and quality assessment

(Interviewer) "The next set of questions will focus on your experiences of learning how to do systematic reviews".

C1. How did you learn to do systematic reviews?
If the interviewee mentions 'training' go to question C2
If they don't mention training, but report through 'practical experience, literature, distance learning, or other' go straight to question C3

C2. Can you tell me a bit more about the training you received?
- *(prompt,* e.g. what format did it take?)
- *(probe where appropriate)* Can you tell me more about this?

C3. How did this help you, if at all?
- *(probe where appropriate)* Can you tell me more about this?

C4. Do you think it adequately prepared you for systematic reviewing?
- *(probe where appropriate)* Why?

C5. Do you have any suggestions for improving the way in which people learn to do systematic reviews?
- *(prompt,* e.g. on-going training during a review?)
- *(probe where appropriate)* Can you tell me more about this?

D Providing training for systematic reviewing

(Note. to interviewer – this section to be covered only with interviewees who have provided support and training for systematic reviews)

I'd like to ask you some questions about your role in providing training and support for systematic reviews

D1. Can you tell me a bit about the training you provide?
- *(probe where appropriate)* Can you tell me more about this?

D2. Which aspects of systematic reviewing does your training cover?
- *(probe where appropriate)* Can you tell me more about this?

D3. Are there any issues that people tend to find difficult to get to grips with?
- *(probe where appropriate)* Why?

D4. Which learning methods do you think are most effective for systematic reviews?
- *(probe where appropriate)* Can you tell me more about this?

D5. Do you think the training currently available adequately addresses the issues most relevant to health promotion?
- *(probe where appropriate)* Why?

D6. Do you have any suggestions for improving the way in which people are trained and supported to do systematic reviews /assess quality? *(NB. Don't ask this question if the interviewee has already answered Question C5*

E – Methods of assessing quality

(Interviewer) "This next set of questions deal with how study quality is assessed in systematic reviews".

E1 Do you routinely assess the quality of the studies in your reviews?
(if 'yes' go to question E2, if 'no' go to question E6)

E2. Can you tell me a bit about how quality is assessed?
- *(prompt,* e.g. how many people are involved? At what stage of the review is it done?)
- *(probe where appropriate)* Can you tell me more about this?

E3. Could you describe the instrument / criteria you use to assess quality?
- *(prompt,* e.g. is it a scale, or checklist?)
- *(probe where appropriate)* Can you tell me more about this?

E4. Why did you choose this instrument / criteria?
- *(prompt,* e.g. it is validated and recommended?)
- *(probe where appropriate)* Can you tell me more about this?

E5. What factors, in your experience, makes quality assessment easier to do?
- *(Prompt,* because you've had good training in critical appraisal?)
- *(probe where appropriate)* Can you tell me more about this?

(go to question E7)

E6. Are there any reasons why you do not assess quality in your systematic reviews?
- *(probe where appropriate)* Can you tell me more about this?

E7 Do you think there is any consensus on what criteria should be used to assess the quality of studies in health promotion?
(probe)
- Why?
- If 'no' Which are the areas of greatest disagreement?

F - The future for systematic reviews

F1. Do you have any suggestions for other issues related to quality that systematic reviews in health promotion should be taking into account?

- *(probe where appropriate)* Can you tell me more about this?

F2. Are there any questions you would like to ask me?

F3. Is there anything you would like to add to what you've said already?

Notes to interviewer:

- Tell them that this concludes the interview.
- Before finishing ask the interviewee if they would like to see a copy of the interview transcript
- Thank them for their time

Appendix 6 – Sampling frame: key characteristics of Cochrane health promotion / public health reviews (n=145)

Table 6.1 - Breakdown of country of corresponding author of health promotion and public health Cochrane systematic reviews (n=145 reviews)

UK	68	Germany	1
USA	15	Netherlands	1
Canada	12	Thailand	1
Australia	20	Switzerland	2
New Zealand	1	Spain	4
China	2	Nigeria	1
Denmark	5	Norway	1
Brazil	4	Not reported	3
Italy	4		

Table 6.2 -Breakdown of the topics under systematic review (n=145 reviews)

Accidents	1	Immunisation uptake	1
Alcohol	4	Injury prevention / safety	22
Asthma + allergy	8	Infectious diseases (non STI)	4
Cardiovascular / coronary heart disease	5	Leisure	0
Cancer	7	Maternal health	8
Child health	1	Medical care	0
Compliance with medication	2	Mental health	8
Congenital conditions	1	Obesity	3
Delinquency	1	Oral health	5
Diabetes	0	Parenting	0
Disability	1	Physical abuse	0
Disease	0	Physical activity	3
Drugs	2	Pregnancy prevention / contraception	2
Eating disorder	0	Problem behaviour	0
Education system	0	Screening and screening uptake	7
Emotional abuse	0	Sexual abuse	0

Epilepsy	0	Sexual health	1
Eye disease	4	Solvents	0
Health promotion in general	4	STI	7
Healthy eating	0	Suicide	0
Hygiene	0	Tobacco	30
Inequalities	0	Workplace	0

NB. Total numbers exceed 145 as reviews could feature more than one topic

Table 6.3 - Breakdown of the types of intervention reviewed (n=145 reviews)

Physical activity	13
Advice	20
Bio-feedback	7
Counselling and psychological interventions	44
Education	60
Environmental modification	21
Immunisation and uptake	4
Incentives	7
Legislation	8
Parent training	6
Physical activity	13
Professional training	6
Regulation	13
Rehabilitation	2
Resource access	15
Risk assessment	11
Screening and testing	17
Service access	6
Skill development	29
Social support	8
Treatment	5
Medication / supplementation	47
Device	17

NB. Total numbers exceed 145 as reviews could feature more than one type of intervention

Appendix 7 – Strategies to recruit interviewees for Stage 2 of the research

As mentioned in Section 5.4.3 of this thesis, four different recruitment strategies were implemented. These are described and justified below.

Strategy 1 – Direct email to Cochrane systematic reviewers

A representative sample of up to 10% of the 145 Cochrane health promotion / public health reviews was taken, and the lead authors were approached directly, via email. As some of the details of the scientific program had been published prior to the conference (via the conference website), I was aware which authors in the sample would be attending to give presentations and potentially would be available to be interviewed.

Strategy 2 – Email from Cochrane Review Group Co-ordinators

The Review Group Coordinators (RGCs) of seven of the 50 Cochrane Review Groups (CRGs) were contacted to ask if they would kindly forward my invitation email to eligible systematic reviewers in their groups. Between them these groups contributed half of the 145 health promotion / public health reviews on the Cochrane library.

Strategy 3 – Email from the Cochrane Health Promotion and Public Health Field

Colleagues at the Cochrane Health Promotion and Public Health Field kindly emailed all the members of their contact database on my behalf, inviting volunteers to contact me directly. Their contact database includes systematic reviewers from all over the world, and was considered to have great potential as a means of accessing an international sample.

Strategy 4 – Snowballing and opportunistic sampling

The fourth strategy occurred as a consequence of strategies one to three. Existing interviewees were asked if they could recommend anyone else relevant that I could approach for possible inclusion in the study. This is a process known as ‘snowball sampling’ and is recognised as a pragmatic way of recruiting people into a research study (Bowling, 2002).

References

Acheson, D. 1998, *Independent Inquiry into Inequalities in Health: Report*, HMSO, London.

Agency for Health Care Policy and Research 1998, *Evidence-Based Practice Centers: Overview.* Rockville MD: AHCRQ.

Aldana, S. G. & Pronk, N. P. 2001, "Health promotion programs, modifiable health risks, and employee absenteeism", *Journal of Occupational and Environmental Medicine*, vol. 43, no. 1, pp. 36-46.

Allen, R. & Stockley, N. 2008, *NIHR Health Technology Assessment Programme Annual Report 2007*. Southampton: University of Southampton.

Altman, D. 2000, *Statistics with Confidence : Confidence Intervals and Statistical Guidelines (2nd edition).* London: BMJ books.

Altman, D. G. 2005, "Endorsement of the CONSORT statement by high impact medical journals: survey of instructions for authors", *British Medical Journal*, vol. 330, pp. 1056-1057.

Altman, D. G. & Bland, J. M. 1999, "Statistics notes. Treatment allocation in controlled trials: why randomise?", *BMJ*, vol. 318, no. 7192, p. 1209.

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P. C., & Lang, T. 2001, "The revised CONSORT statement for reporting randomized trials: explanation and elaboration", *Annals of Internal Medicine*, vol. 134, no. 8, pp. 663-694.

Anderson, L. M., Fielding, J. E., Fullilove, M. T., Scrimshaw, S. C., Carande-Kulis, V. G., & Task Force on Community Preventive Services 2003, "Methods for conducting systematic reviews of the evidence of effectiveness and economic efficiency of interventions to promote healthy social environments", *American Journal of Preventive Medicine*, vol. 24, no. 3:Suppl, p. Suppl-31.

Armstrong, R., Waters, E., & Doyle, J. 2008, "Chapter 21: Reviews in health promotion and public health," in *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 (updated February 2008)*, J. Higgins & S. Green, eds. Oxford: The Cochrane Collaboration.

Bailey, K. 1994, *Methods of Social Research*, 4th edition, The Free Press, New York.

Bambra, C., Whitehead, M., & Hamilton, V. 2005, "Does 'welfare-to-work' work? A systematic review of the effectiveness of the UK's welfare-to-work programmes for people with a disability or chronic illness.", *Social Science & Medicine*, vol. 60, no. 9, pp. 1905-1918.

Baranowski, T., Simons-Morton, B., Hooks, P., Henske, J., Tiernan, K., Dunn, J. K., Burkhalter, H., Harper, J., & Palmer, J. 1990, "A center-based program for exercise change among black-American families", *Health Education Quarterly.*, vol. 17, no. 2, pp. 179-196.

Bartlett, C., Doyal, L., Ebrahim, S., Davey, P., Bachmann, M., Egger, M., & Dieppe, P. /1/15, "The causes and effects of socio-demographic exclusions from clinical trials.", *Health Technology Assessment*, vol. 9, no. 38.

Bazeley, P. 2003, "Computerized Data Analysis for Mixed Methods Research," in *Handbook of Mixed Methods in Social & Behavioural Research*, A. Tashakkori & C. Teddlie, eds., Sage, Thousand Oaks, California, pp. 385-422.

Bell, S. G., Newcomer, S. F., Bachrach, C., Borawski, E., Jemmott, J. B., III, Morrison, D., Stanton, B., Tortolero, S., & Zimmerman, R. 2007, "Challenges in replicating interventions", *Journal of Adolescent Health*, vol. 40, no. 6, pp. 514-520.

Biglan, A., Severson, H., Ary, D., Faller, C., Gallison, C., Thompson, R., Glasgow, R., & Lichtenstein, E. 1987, "Do smoking prevention programs really work? Attrition and the internal and external validity of an evaluation of a refusal skills training program", *Journal of Behavioral Medicine,* vol. 10, no. 2, pp. 159-71.

Black, N., Murphy, M., Lamping, D., McKee, M., Sanderson, C., Askham, J., & Marteau, T. 1999, "Consensus development methods: a review of best practice in creating clinical guidelines.", *Journal of Health Services & Research Policy*, vol. 4, no. 4, pp. 236-248.

Black, N. 1996, "Why we need observational studies to evaluate the effectiveness of health care", *British Medical Journal*, vol. 312, no. 7040, pp. 1215-1218.

Bonell, C., Oakley, A., Hargreaves, J., Strange, V., & Rees, R. 2006, "Assessment of generalisability in trials of health interventions: suggested framework and systematic review.", *British Medical Journal*, vol. 333, no. 7563, pp. 346-349.

Bonell, C. & Imrie, J. 2001, "Behavioural interventions to prevent HIV infection: rapid evolution, increasing rigour, moderate success", *British Medical Bulletin*, vol. 58, pp. 155-170.

Booth RE & Watters JK 1994, "How effective are risk-reduction interventions targeting injecting drug users?", *AIDS*, vol. 8, pp. 1515-1524.

Boruch, R., Soydan, H., & de Moya, D. 2004, "The Campbell Collaboration", *Brief Treatment and Crisis Intervention*, vol. 4, no. 3, pp. 277-287.

Boutron, I., Guittet, L., Estellat, C., Moher, D., Hrobjartsson, A., & Ravaud, P. 2007, "Reporting methods of blinding in randomized trials assessing nonpharmacological treatments.", *PLoS Medicine / Public Library of Science*, vol. 4, no. 2, p. e61.

Boutron, I., Moher, D., Altman, D. G., Schulz, K. F., Ravaud, P., & CONSORT Group 2008, "Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration", *Annals of Internal Medicine*, vol. 148, no. 4, pp. 295-309.

Bowling, A. 2002, *Research Methods in Health.* Maidenhead: Open University Press.

Briss, P. A., Rodewald, L. E., Hinman, A. R., Shefer, A. M., Strikas, R. A., Bernier, R. R., Carande-Kulis, V. G., Yusuf, H. R., Ndiaye, S. M., & Williams, S. M. 2000, "Reviews of evidence regarding interventions to improve vaccination coverage in children, adolescents, and adults. The Task Force on Community Preventive Services.", *American Journal of Preventive Medicine*, vol. 18, no. 1:Suppl, pp. 97-140.

Britton, A., McKee, M., Black, N., McPherson, K., Sanderson, C., & Bain, C. 1998, "Choosing between randomised and non-randomised studies: a systematic review", *Health Technology Assessment*, vol. 2, no. 13.

Broom, A. & Willis, E. 2007, "Competing Paradigms in Health Research," in *Researching Health: Qualitative, Quantitative and Mixed Methods*, M. Saks & J. Allsop, eds., Sage, London, pp. 17-31.

Brown, H. 2007, "How impact factors changed medical publishing--and science", *British Medical Journal, 334(7593):561-4*.

Boyd, N. & Windsor, R. 1993, "A meta-evaluation of nutrition education intervention research among pregnant women", *Health Education Quarterly*, vol. 20, no. 3, pp. 327-345.

Brunton, G., Waters, E., Doyle, J., Kavanagh, J., Shepherd, J., Rees, R., Harden, A., Oliver, S., & Oakley, A. "Finding a haystack among the needles: health promotion and public health reviews in the Cochrane Database of systematic reviews (abstract)" - Poster presentation at the *X Cochrane Colloquium*, Stavanger, Norway, 31st July - 3rd August 2002.

Brunton, G., Thomas, J., Harden, A., Rees, R., Kavanagh, J., Oliver, S., Shepherd, J., & Oakley, A. 2005, "Promoting physical activity amongst children outside of physical education classes: a systematic review integrating intervention studies and qualitative studies", *Health Education Journal*, vol. 64, no. 4.

Bryman, A. 2006, "Integrating quantitative and qualitative research: how is it done?", *Qualitative Research*, vol. 6, pp. 97-113.

Bryman, A. 2008, "The End of the Paradigm Wars?," in *The SAGE Handbook of Social Research Methods*, P. Alasuutari, L. Bickman, & J. Brannen, eds., Sage, London, pp. 13-25.

Campbell, D. T. & Stanley, J. C. 1963, *Experimentation and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Campbell, D. T. 1986, "Relabeling internal and external validity for applied social scientists," in *Advances in Quasi-Experimental Design and Analysis: New Directions for Program Evaluation*, W. Trochim, ed. San Fransisco: Jossey-Bass.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A. L., Sandercock, P., Spiegelhalter, D., & Tyrer, P. 2000, "Framework for design and evaluation of complex interventions to improve health", *British Medical Journal*, vol. 321, no. 7262, pp. 694-696.

Campbell, M. 2000, "A systematic review of the effectiveness of environmental awareness interventions", *Canadian Journal of Public Health*, vol. 91, no. 2, pp. 137-143.

Campbell, N. C., Murray, E., Darbyshire, J., Emery, J., Farmer, A., Griffiths, F., Guthrie, B., Lester, H., Wilson, P., & Kinmonth, A. L. 2007, "Designing and evaluating complex interventions to improve health care.", *British Medical Journal, 334(7591):455-9.*

Carter, S. & Henderson, L. 2005, "Approaches to qualitative data collection in social science," in *Handbook of Health Research Methods: Investigation, Measurement and Analysis*, A. Bowling & S. Ebrahim, eds. Maidenhead: Open University Press. pp. 215-229.

Catford, J. 1993, "Auditing health promotion: what are the vital signs of quality?", *Health Promotion International*, vol. 8, pp. 67-68.

Centre for Reviews and Dissemination 2009, *Systematic Reviews: CRD's guidance for Undertaking Reviews in Health Care*. York: CRD.

Centre for Reviews and Dissemination 2001, *Undertaking Systematic Reviews of Research on Effectiveness: CRD's Guidance for those Carrying Out or Commissioning Reviews (2nd Edition)*. York: York Publishing Services Ltd.

Chalmers, T. C., Celano, P., Sacks, H. S., & Smith, H., Jr. 1983, "Bias in treatment assignment in controlled clinical trials", *New England Journal of Medicine.*, vol. 309, no. 22, pp. 1358-1361.

Ciliska, D., Miles, E., OBrien, M. A., Turl, C., Tomasik, H. H., Donovan, U., & Beyers, J. 2000, "Effectiveness of community-based interventions to increase fruit and vegetable consumption", *Journal of Nutrition Education*, vol. 32, no. 6, pp. 341-352.

Clarke, M. & Oxman, A. 2001, *Cochrane Reviewers Handbook Glossary 4.1.4 (Updated October 2001)*, Oxford: Update Software.

Clarke, M. 2006, "The Cochrane Collaboration," in *Public Health Evidence: Tackling Health Inequalities*, A. Killoran, C. Swann, & M. Kelly, eds. Oxford: Oxford University Press, pp. 115-124.

Cochrane, A. L. 1979, "1931-1971, a critical review, with particular reference to the medical profession," in *Medicines for the Year 2000*. London: Office of Health Economics, pp. 1-11.

Cochrane Eyes and Vision Group. 2006. *Cochrane Eyes and Vision Group Review Development Guidelines*. http://www.cochraneeyes.org/documents/CEVGRevDevGuide_Feb2006.pdf [accessed 23/4/09]

Cochrane Collaboration. 2004. *The Cochrane Logo Resources Pages*. http://www.cochrane.org/logo/ [accessed 24/4/09]

Cochrane Collaboration. 2008. *The Cochrane Library* http://www3.interscience.wiley.com/cgi-bin/mrwhome/106568753/HOME [accessed 24/10/08]

Cochrane Collaboration. 2005. *Cochrane Colloquium*. http://www.colloquium.info [accessed 5/5/05]

Cochrane Consumer Network. 2008. *Consumers in Cochrane*. http://www.cochrane.org/consumers/cinc.htm [accessed 3/8/08]

Cohen, J. 1988, *Statistical Power Analysis for the Behavioural Sciences*, (2nd edition). Hillsdale, NJ: Erlbaum.

Cohen, L., Manion, L., & Morrison, K. 2007, *Research Methods in Education* (6th edition) London: Routledge.

Colditz, G. A., Miller, J. N., & Mosteller, F. 1989, "How study design affects outcomes in comparisons of therapy. I: Medical", *Statistics in Medicine.8(4):441-54*.

Cook, T. D. & Campbell, D. T. 1979, *Quasi-Experimentation: Design and Analysis Issues for Field Settings,* Chicago: Rand McNally.

Cooper, H. 1989, *Integrating Research* (2nd edition). California: Sage.

Cooper, H. & Hedges, L. 1994, *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Coren, E. & Fisher, M. 2006, *The Conduct of Systematic Research Reviews for SCIE Knowledge Reviews*. London: Social Care Institute for Excellence.

Coren, E. & Bates, S. 2006, "Systematic mapping - a new development in the evidence base for social care [abstract]", *XIV Cochrane Colloquium. 23rd to 26th October, Dublin., Ireland.*

Coyle, K. K., Kirby, D. B., Robin, L. E., Banspach, S. W., Baumler, E., & Glassman, J. R. 2006, "All4You! A randomized trial of an HIV, other STDs, and pregnancy prevention intervention for alternative school students", *AIDS Education and Prevention*, vol. 18, pp. 187-203.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. 2008, *Developing and Evaluating Complex Interventions*. London: Medical Research Council.

Cross, J. E., Saunders, C. M., & Bartelli, D. 1998, "The Effectiveness of Educational and Needle Exchange Programs: A Meta-Analysis of HIV Prevention Strategies for Injecting Drug Users", *Quality and Quantity*, vol. 32, pp. 165-180.

Cross-Government Obesity Unit 2008, *Healthy Weight, Healthy Lives: a Cross-Government Strategy for England*. London: Cross-Government Obesity Unit.

Crowley, P. 2000, "Prophylactic corticosteroids for preterm birth.", *Cochrane Database of Systematic Reviews.*(2):CD000065

Cullum, N. 2008, *Evidence-based Nursing: An Introduction.* Oxford: Wiley-Blackwell.

Dalkey, N. & Helmer, O. 1963, "An experimental application of the Delphi method to the use of experts", *Management Science*, vol. 9, no. 3, pp. 458-467.

Davis, P. & Scott, A. 2007, "Health research sampling methods," in *Researching Health: Qualitative, Quantitative and Mixed Methods*, M. Saks & J. Allsop, eds., Sage, London, pp. 155-173.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., Petticrew, M., Altman, D. G., International Stroke Trial Collaborative Group, & European Carotid Surgery Trial Collaborative Group 2003, "Evaluating non-randomised intervention studies.", *Health Technology Assessment*, vol. 7, no. 27.

Delbecq, A. & Van de Ven, A. 1971, "A group process model for problem identification and program planning", *Journal of Applied Behavioural Sciences*, vol. 7, no. 4, pp. 467-492.

Denzin, N. K. 1978, *The Research Act in Sociology: A Theoretical Introduction to Sociological Methods.* New York: McGraw-Hill.

Department of Health 2001, *The National Strategy for Sexual Health and HIV*, HMSO, London.

Department of Health 2002a, *National Suicide Prevention Strategy for England*, London: Department of Health.

Department of Health 2002b, *The National Strategy for Sexual Health and HIV: implementation and action plan*. London: HMSO.

Department of Health 2003, *Directions to Primary Care Trusts and NHS Trusts in England Concerning Arrangements for the Funding of Technology Appraisal Guidance from the National Institute for Clinical Excellence (NICE)*. London: Department of Health.

Department of Health 2004, *Choosing Health: Making Healthy Choices Easier*. London: HMSO.

Department of Health 2006, *Best Research for Best Health - A New National Health Research Strategy*. London: HMSO.

Department of Health 2009, *The Coronary Heart Disease National Service Framework: Building on Excellence, Maintaining Progress. Progress report for 2008*. London: Department of Health.

Des, J., Lyles, C., Crepaz, N., & TREND Group 2004, "Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement", *American Journal of Public Health*, vol. 94, no. 3, pp. 361-366.

Detsky, A., Naylor, C., O'Rourke, K., McGeer, A., & L'Abbe, A. 1992, "Incorporating variations in the quality of individual randomized trials into meta-analysis", *Journal of Clinical Epidemiology*, vol. 45, no. 3, pp. 255-265.

Devereaux, P. J., Manns, B. J., Ghali, W. A., Quan, H., & Guyatt, G. H. 2002, "The reporting of methodological factors in randomized controlled trials and the association with a journal policy to promote adherence to the Consolidated Standards of Reporting Trials (CONSORT) checklist", *Controlled Clinical Trials*, vol. 23, pp. 380-388.

DiCenso, A., Guyatt, G., & Ciliska, D. 2004, *Evidence-Based Nursing: A Guide to Clinical Practice.* London: Mosby.

DiCenso, A., Guyatt, G., Willan, A., & Griffith, L. 2002, "Interventions to reduce unintended pregnancies among adolescents: systematic review of randomised controlled trials.[comment].", *British Medical Journal*, vol. 324, no. 7351, p. 1426.

Dickersin, K. 1997, "How important is publication bias? a synthesis of available data", *AIDS Education and Prevention* no. 9(suppl A), pp. 15-21.

Dickersin, K. & Berlin, J. A. 1992, "Meta-analysis: state of the science.", *Epidemiological Review*, vol. 14, pp. 154-176.

Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., & Smith, H., Jr. 1987, "Publication bias and clinical trials", *Controlled Clinical Trials.*, vol. 8, no. 4, pp. 343-353.

Dishman RK & Buckworth J 1996, "Increasing physical activity: a quantitative synthesis", *Medicine & Science in Sports & Exercise*, vol. 28, no. 6, pp. 706-719.

Dixon-Woods, M. & Fitzpatrick, R. 2001, "Qualitative research in systematic reviews (editorial)", *British Medical Journal*, vol. 323, pp. 765-766.

Donnelly, A., Snowden, H. M., Renfrew, M. J., & Woolridge, M. W. 2000, "Commercial hospital discharge packs for breastfeeding women.", *Cochrane Database of Systematic Reviews.* (2):CD002075.

Doyle, J., Waters, E., Yach, D., McQueen, D., De, F. A., Stewart, T., Reddy, P., Gulmezoglu, A. M., Galea, G., & Portela, A. 2005, "Global priority setting for Cochrane systematic reviews

of health promotion and public health research", *Journal of Epidemiology & Community Health*, vol. 59, no. 3, pp. 193-197.

Drummond, M., Schulpher, M., Torrance, G., O'Brien, B., & Stoddart, G. 2005, *Methods for the Economic Evaluation of Health Care Programmes.* (2nd edition). Oxford: Oxford University Press.

Dumas, J. E., Lynch, A. M., Laughlin, J. E., Phillips, S. E., & Prinz, R. J. 2001, "Promoting intervention fidelity. Conceptual issues, methods, and preliminary results from the EARLY ALLIANCE prevention trial", *American Journal of Preventive Medicine.*, vol. 20, no. 1:Suppl, p. Suppl-47.

Dunn, A. L., Andersen, R. E., & Jakicic, J. M. 1998, "Lifestyle physical activity interventions. History, short- and long-term effects, and recommendations", *American Journal of Preventive Medicine*, vol. 15, no. 4, pp. 398-412.

Dyson, L., McCormick, F., & Renfrew, M. J. 2005, "Interventions for promoting the initiation of breastfeeding.", *Cochrane Database of Systematic Reviews* no. 2, CD001688.

Egger, M., Davey Smith, G., & Schneider, M. 2001, "Systematic reviews of observational studies," in *Systematic Reviews in Health Care: Meta Analysis in Context*, M. Egger, G. Davey Smith, & D. Altman, eds., London: BMJ Books.

Egger, M., Juni, P., Bartlett, C., Holenstein, F., & Sterne, J. 2003, "How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study.", *Health Technology Assessment,* vol.7, no. 1.

Egger, M., Juni, P., & Bartlett, C. 2001, "Value of flow diagrams in reports of randomized controlled trials", *Journal of the American Medical Association*, vol. 285, no. 15, pp. 1996-1999.

Elbourne, D. R. & Campbell, M. K. 2001, "Extending the CONSORT statement to cluster randomized trials: for discussion", *Statistics in Medicine*, vol. 20, no. 3, pp. 489-496

Elford, J., Sherr, L., Bolding, G., Serle, F., & Maguire, M. 2002, "Peer-led HIV prevention among gay men in London: process evaluation", *AIDS Care. 14(3):351-60.*

Ellis, S., Barnett-Page, E., Morgan, A., Taylor, L., Walters, R., & Goodrick, J. 2003, *HIV Prevention: a Review of Reviews Assessing the Effectiveness of Intervention to Reduce the Risk of Sexual Transmission: Evidence Briefing*. London: Health Development Agency.

Ellis, S. & Grey, A. 2004, *Prevention of Sexually Transmitted Infections (STIs): a Review of Reviews into the Effectiveness of Non-Clinical Interventions*. London: Health Development Agency.

Elwy, A. R., Hart, G. J., Hawkes, S., & Petticrew, M. 2002, "Effectiveness of interventions to prevent sexually transmitted infections and human immunodeficiency virus in heterosexual men: a systematic review.", *Archives of Internal Medicine.*, vol. 162, no. 16, pp. 1818-1830.

Evans, D., Head, D., & Speller, V. 1994, *Assuring Quality in Health Promotion: How to Develop Standards of Good Practice*. London: Health Education Authority.

Fielding, N. & Lee, R. M. 2008, *Computer Analysis and Qualitative Research*. London: Sage..

Fink, A., Kosecoff, J., Chassin, M., & Brook, R. H. 1984, "Consensus methods: characteristics and guidelines for use", *American Journal of Public Health*, vol. 74, no. 9, pp. 979-983.

Fisher, J. D. & Fisher, W. A. 2000, "Theoretical approaches to individual-level change in HIV risk behavior," in *Handbook of HIV prevention*, J. L. Peterson & R. J. DiClemente, eds. New York, NY: Kluwer Academic/Plenum Publishers. xvi, 337 pp.

Fitz-Gibbon, C. 1985, "The implications of meta analysis for educational research", *British Educational Research Journal*, vol. 11, pp. 45-49.

Flay, B. R. 1986, "Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs.", *Preventive Medicine*, vol. 15, no. 5, pp. 451-474.

Fletcher A & Rake C 1998, *Effectiveness of Interventions to Promote Healthy Eating in Elderly People Living in the Community: a Review*, Effectiveness Review nr. 8. London: Health Education Authority.

Foxcroft DR, Lister-Sharp D, & Lowe G 1997, "Alcohol misuse prevention for young people: a systematic review reveals methodological concerns and lack of reliable evidence of effectiveness", *Addiction*, vol. 92, no. 5, pp. 531-537.

Fraser, E. 1996, "How effective are effectiveness reviews?", *Health Education Journal*, vol. 55, pp. 359-362.

Galloe, A. M., Thuesen, L., Kelbaek, H., Thayssen, P., Rasmussen, K., Hansen, P. R., Bligaard, N., Saunamaki, K., Junker, A., Aaroe, J., Abildgaard, U., Ravkilde, J., Engstrom, T., Jensen, J. S., Andersen, H. R., Botker, H. E., Galatius, S., Kristensen, S. D., Madsen, J. K., Krusell, L. R., Abildstrom, S. Z., Stephansen, G. B., Lassen, J. F., & SORT, O. I., I 2008, "Comparison of

paclitaxel- and sirolimus-eluting stents in everyday clinical practice: the SORT OUT II randomized trial.", *Journal of the American Medical Association*, vol. 299, no. 4, pp. 409-416.

General Medical Council 2003, *Tomorrow's Doctors: Recommendations on Undergraduate Medical Education*. London: General Medical Council.

Gibbs, G. 2002, *Qualitative Data Analysis: Explorations with NVivo.* Maidenhead: Open University Press.

Glanz K, Sorensen G, & Farmer A 1996, "The heatlh impact of worksite nutrition and cholesterol intervention programs", *American Journal of Health Promotion*, vol. 10, no. 6, pp. 453-470.

Glaser, B. G. & Strauss, A. L. 1967, *The Discovery of Grounded Theory: Strategies for Qualitative Research.* New York: Gruyter.

Glasziou, P., Irwig, L., Bain, C., & Colditz, G. *Systematic reviews in health care: a practical guide*. Cambridge: Cambridge University Press.

Global Health Equity Group. 2009. *Overview of the Review of Health Inequalities Post 2010 in England (Marmot Review)*. http://www.ucl.ac.uk/gheg/marmotreview [accessed 17/4/09]

Godlee, F. 2006, "Where next for the research assessment exercise?", *British Medical Journal*, vol. 332, no. 7548, p. doi:10.1136/bmj.332.7548.0-f.

Goodare, H. & Lockwood, S. 1999, "Involving patients in clinical research. Improves the quality of research.", *British Medical Journal* , vol. 319, no. 7212, pp. 724-725.

Gorden, R. 1992, *Basic Interviewing Skills.* USA*:* Peacock publishers.

Graham, C. A., Catania, J. A., Brand, R. A., Duong, B., & Canchola, J. A. 2003, "Recalling sexual behavior: A methodological analysis of memory recall bias via interview using the diary as the gold standard", *Journal of Sex Research*, vol. 40, no. 4, pp. 325-332.

Green, J. & Thorogood, N. 2004, *Qualitative Methods for Health Research.* London: Sage publications.

Green, J. 2000, "The role of theory in evidence-based health promotion practice", *Health Education Research*, vol. 15, no. 2, pp. 125-129.

Green, L. W. & Glasgow, R. E. 2006, "Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology.", *Evaluation & the Health Professions*, vol. 29, no. 1, pp. 126-153.

Green, L. & Kreuter, M. 1999, *Health Promotion Planning: an Educational and Ecological Approach (3rd edition).* Boston: McGraw-Hill.

Greenhalgh, T. 1997, "Papers that summarise other papers (systematic reviews and meta-analyses)", *British Medical Journal*, vol. 315, no. 7109, pp. 672-675.

Grimes, D. A. & Schulz, K. F. 2002, "An overview of clinical research: the lay of the land", *Lancet*, vol. 359, no. 9300, pp. 57-61.

Guba, E. G. & Lincoln, Y. S. 2005, "Paradimatic controversies, contradictions, and emerging confluences," in *The Sage Handbook of Qualitative Research*, 3rd edition, N. K. Denzin & Y. S. Lincoln, eds. Thousand Oaks, California: Sage. pp. 191-215.

Guyatt, G. H., DiCenso, A., Farewell, V., Willan, A., & Griffith, L. 2000, "Randomized trials versus observational studies in adolescent pregnancy prevention", *Journal of Clinical Epidemiology.*, vol. 53, no. 2, pp. 167-174.

Gwadz, M. & Rotheram-Borus, M. J. 1992, "Tracking high-risk adolescents longitudinally", *AIDS Education & Prevention*, vol. Suppl, pp. 69-82.

Hallett, T. B., White, P. J., & Garnett, G. P. 2007, "Appropriate evaluation of HIV prevention interventions: from experiment to full-scale implementation.", *Sexually Transmitted Infections*, vol. 83, p. Suppl-60.

Hammersley, M. 2008, "Paradigm war revived? On the diagnosis of resistance to randomized controlled trials and systematic review in education", *International Journal of Research & Method in Education*, vol. 31, no. 1, pp. 3-10.

Hammersley, M. 2006, "Systematic or unsystematic, is that the question? Reflections on the science, art, and politics of reviewing research evidence," in *Public Health Evidence: Changing the Health of the Public*, A. Killoran, C. Swann, & M. Kelly, eds., Oxford University Press, Oxford, pp. 238-249.

Hammersley, M. 2001, "On 'systematic' reviews of research literatures: a 'narrative' response to Evans & Benefield", *British Educational Research Journal*, vol. 27, no. 5, pp. 543-554.

Hammersley, M. 1992, "The paradigm wars: reports from the front", *British Journal of Sociology of Education*, vol. 13, pp. 131-143.

Hammersley, M. & Atkinson, P. 2007, *Ethnography: Principles in Practice* (3rd Edition) London: Routledge.

Hanney, S., Buxton, M., Green, C., Coulson, D., & Raftery, J. 2007, "An assessment of the impact of the NHS Health Technology Assessment Programme.", *Health Technology Assessment*, vol. 11, no. 53.

Harden, A., Garcia, J., Oliver, S., Rees, R., Shepherd, J., Brunton, G., & Oakley, A. 2004, "Applying systematic review methods to studies of people's views: an example from public health research", *Journal of Epidemiology and Community Health*, vol. 58, pp. 794-800.

Harden, A. 2001, "Finding research evidence: systematic searching," in *Using Research for Effective Health Promotion*, S. Oliver & G. Peersman, eds. Buckingham: Open University Press, pp. 47-68.

Harden, A., Peersman, G., Oliver, S., & Oakley, A. 1999, "Identifying relevant primary research on electronic databases to inform decision-making in health promotion; the case of sexual health promotion", *Health Education Journal*, vol. 58, no. 3, pp. 290-301.

Harden, A. & Thomas, J. 2005, "Methodological issues in combining diverse study types in systematic reviews", *Journal of Social Research Methodology: Theory and Practice*, vol. 8, pp. 257-271.

Harden, A. & Oliver, S. 2001, "Who's listening? Systematically reviewing for ethics and empowerment," in *Using Research for Effective Health Promotion*, S. Oliver & G. Peersman, eds. Buckingham: Open University Press. pp. 123-137.

Hawe, P., Shiell, A., & Riley, T. 2004, "Complex interventions: how "out of control" can a randomised controlled trial be?", *British Medical Journal*, vol. 328, no. 7455, pp. 1561-1563.

Hawkins, S. S. & Law, C. 2005, "Patterns of research activity related to government policy: a UK web based survey.", *Archives of Disease in Childhood*, vol. 90, no. 11, pp. 1107-1111.

Health Protection Agency 2008, *Sexually Transmitted Infections and Young People in the United Kingdom*. London: Health Protection Agency, Centre for Infections.

Herbert, R. D. & Bo, K. 2005, "Analysis of quality of interventions in systematic reviews", *British Medical Journal*, vol. 331, no. 7515, pp. 507-509.

Higgins, J. P. & Green, S. 2008, *Cochrane Handbook for Systematic Reviews of Interventions (Version 5.0) updated February 2008*. Oxford: The Cochrane Collaboration.

Hillsdon, M., Foster, C., & Thorogood, M. 2005, "Interventions for promoting physical activity.", *Cochrane Database of Systematic Reviews* no. 1, p. CD003180.

Huibers, M. J., Beurskens, A. J., Bleijenberg, G., & van Schayck, C. P. 2003, "The effectiveness of psychosocial interventions delivered by general practitioners.", *Cochrane Database of Systematic Reviews* no. 2, p. CD003494.

Hursti UK & Sjoden P 1997, "Changing food habits in children and adolescents: Experiences from intervention studies", *Scandinavian Journal of Nutrition*, vol. 41, pp. 102-110.

Hutchings, A. & Raine, R. 2006, "A systematic review of factors affecting the judgements produced by formal consensus development methods in health care", *Journal of Health Services & Research Policy*, vol. 11, no. 3, pp. 172-179.

Jackson, N., Waters, E., & Guidelines for Systematic Reviews in Health Promotion and Public Health Taskforce 2005, "Criteria for the systematic review of health promotion and public health interventions", *Health Promotion International*, vol. 20, no. 4, pp. 367-374.

Jackson, N., Waters, E., & Guidelines for Systematic Reviews of Health Promotion and Public Health Interventions Taskforce 2004, "The challenges of systematically reviewing public health interventions", *Journal of Public Health*, vol. 26, no. 3, pp. 303-307.

Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J., Gavaghan, D. J., & McQuay, H. J. 1996, "Assessing the quality of reports of randomized clinical trials: is blinding necessary?", *Controlled Clinical Trials*, vol. 17, no. 1, pp. 1-12.

Joint Committee on Standards for Educational Evaluation 1994, *The Program Evaluation Standards (2nd Edition)*, Sage, Newbury Park, California.

Jones, J., Shepherd, J., Baxter, L., Gospodarevskaya, E., Hartwell, D., Harris, P., & Price, A. 2009, "Adefovir dipivoxil and pegylated interferon alfa for the treatment of chronic hepatitis B - an updated systematic review and economic evaluation", *Health Technology Assessment (in press)*.

Jones, J. & Hunter, D. 1995, "Qualitative Research: Consensus methods for medical and health services research", *British Medical Journal*, vol. 311, no. 7001, pp. 376-380.

Juni, P., Altman, D. G., & Egger, M. 2001, "Systematic reviews in health care: Assessing the quality of controlled clinical trials", *British Medical Journal*, vol. 323, no. 7303, pp. 42-46.

Juni, P., Witschi, A., Bloch, R., & Egger, M. 1999, "The hazards of scoring the quality of clinical trials for meta-analysis", *Journal of the American Medical Association*, vol. 282, no. 11, pp. 1054-1060.

Kamali, A., Quigley, M., Nakiyingi, J., Kinsman, J., Kengeya-Kayondo, J., Gopal, R., Ojwiya, A., Hughes, P., Carpenter, L. M., & Whitworth, J. 2003, "Syndromic management of sexually-transmitted infections and behaviour change interventions on transmission of HIV-1 in rural Uganda: a community randomised trial", *Lancet*, vol. 361, no. 9358, pp. 645-652.

Katz, D. L., Williams, A. L., Girard, C., Goodman, J., Comerford, B., Behrman, A., & Bracken, M. B. 2003, "The evidence base for complementary and alternative medicine: methods of Evidence Mapping with application to CAM", *Alternative Therapies in Health & Medicine*, vol. 9, no. 4, pp. 22-30.

Keogh-Brown, M. R., Bachmann, M. O., Shepstone, L., Hewitt, C., Howe, A., Ramsay, C. R., Song, F., Miles, J. N., Torgerson, D. J., Miles, S., Elbourne, D., Harvey, I., & Campbell, M. J. 2007, "Contamination in trials of educational interventions", *Health Technology Assessment*, vol. 11, no. 43.

Kelly, M. P., McDaid, D., Ludbrook, A., & Powell, J. 2005, *Economic Appraisal of Public Health Interventions*, Health Development Agency, London.

Kelly, M. P. 2006a, "The development of an evidence-based approach to tackling health inequalities in Britain," in *Public Health Evidence: Tackling Health Inequalities*, A. Killoran, C. Swann, & M. Kelly, eds. Oxford: Oxford University Press. pp. 41-62.

Kelly, M. P. 2006b, "New NICE and public health", *Health Education*, vol. 106, no. 3, pp. 181-184.

Kelly, M. P. 2005, "Public health guidance and the role of new NICE", *Public Health*, vol. 119, no. 11, pp. 960-968.

Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. 2003, *Systematic Reviews to Support Evidence-Based Medicine: How to Review and Apply Findings of Healthcare Research.* Oxford: The Royal Society of Medicine Press.

Killip, S., Mahfoud, Z., & Pearce, K. 2004, "What is an intracluster correlation coefficient? Crucial concepts for primary care researchers", *Annals of Family Medicine*, vol. 2, no. 3, pp. 204-208.

Kirby, D., Short, L., Collings, J., Rugg, D., Kolbe, L., Howard, M., Miller, B., Sonenstein, F., & Zabin, L. 1994, "School-based programs to reduce sexual risk behaviors: a review of effectiveness", *Public Health Reports*, vol. 109, no. 3, pp. 339-360.

Kirby, D., Laris, B. A., & Rolleri, L. 2006, *The Impact of Sex and HIV Education Programs in Schools and Communities on Sexual Behaviors among Young Adults*. North Carolina: Family Health International, YouthNet program.

Kjaergard, L. L., Nikolova, D., & Gluud, C. 1999, "Randomized clinical trials in Hepatology: predictors of quality", *Hepatology*, vol. 30, no. 5, pp. 1334-1338.

Kjaergard, L. L., Villumsen, J., & Gluud, C. 2001, "Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses", *Annals of Internal Medicine*, vol. 135, no. 11, pp. 982-989.

Kleijnen, J., Gotzsche, P., Kunz, R., Oxman, A., & Chalmers, I. 1997, "So what's so special about randomization?," in *Non-Random Reflections on Health Services Research*, A. Maynard & I. Chalmers, eds. London: BMJ Publishing Group.

Kunz, R. & Oxman, A. D. 1998, "The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials", *British Medical Journal*, vol. 317, no. 7167, pp. 1185-1190.

Kvale, S. 1996, *Interviews : an introduction to qualitative research interviewing.* California: Thousand Oaks.

Learmonth, A. & Watson, N. 1999, "Constructing evidence based health promotion: perspectives from the field", *Critical Public Health*, vol. 19, pp. 317-333.

Leizorovicz, A., Haugh, M. C., Chapuis, F. R., Samama, M. M., & Boissel, J. P. 1992, "Low molecular weight heparin in prevention of perioperative thrombosis", *British Medical Journal*, vol. 305, no. 6859, pp. 913-920.

Lexchin, J., Bero, L. A., Djulbegovic, B., & Clark, O. 2003, "Pharmaceutical industry sponsorship and research outcome and quality: systematic review", *British Medical Journal,* vol. 326, no. 7400, pp. 1167-70.

Lilford, R. J., Richardson, A., Stevens, A., Fitzpatrick, R., Edwards, S., Rock, F., & Hutton, J. L. 2001, "Issues in methodological research: perspectives from researchers and commissioners", *Health Technology Assessment*, vol. 5, no. 8.

Lingard, L., Albert, M., & Levinson, W. 2008, "Grounded theory, mixed methods, and action research", *British Medical Journal*, vol. 337, no. 7667, p. a567.

Littlejohns, P. 2006, "The National Institute for Health and Clinical Excellence," in *Public Health Evidence: Tackling Health Inequalities*, A. Killoran, C. Swann, & M. Kelly, eds. Oxford: Oxford University Press, pp. 167-186.

Liu, J. & Yang, M. 2003, "How trial reports of alternative therapies comply with the CONSORT statement? A literature review on Chinese medicines [abstract]", *XI Cochrane Colloquium.: Evidence., Health Care and Culture. Oct.26th to 31st, Barcelona., Spain.*

Loevinsohn, B. P. 1990, "Health education interventions in developing countries: a methodological review of published articles", *International Journal of Epidemiology.*, vol. 19, no. 4, pp. 788-794.

Low, J. 2007, "Unstructured interviews and health research," in *Researching Health: Qualitative, Quantitative and Mixed Methods*, M. Saks & J. Allsop, eds., Sage, London, pp. 74-91.

Lumley, J., Oliver, S., & Waters, E. 2000, "Interventions for promoting smoking cessation during pregnancy", *Cochrane Database of Systematic Reviews.* no. 2, p. CD001055.

Macdonald, T. 1998, *Rethinking Health Promotion: A Global Approach.* London: Routledge.

Macdonald, G. 1997, "Social work: beyond control?," in *Non-Random Reflections on Health Services Research*, A. Maynard & I. Chalmers, eds. London: BMJ Publishing Group.

Macdonald, G., Sheldon, B., & Gillespie, J. 1992, "Contemporary-Studies of the Effectiveness of Social-Work", *British Journal of Social Work*, vol. 22, no. 6, pp. 615-643.

Maynard, A. & Chalmers, I. 1997, *Non-Random Reflections on Health Services Research.* London: BMJ Publishing Group.

McBurney, D. H. 2001, *Research Methods.* Belmont, California: Wadsworth/Thomson Learning.

McGuire, C. 2006, "Building the Evidence Base: the Contribution of the Department of Health's Policy Research Programme (England)," in *Public Health Evidence: Tackling Health Inequalities*, A. Killoran, C. Swann, & M. Kelly, eds., Oxford University Press, Oxford, pp. 83-94.

McMichael, C., Waters, E., & Volmink, J. 2005, "Evidence-based public health: what does it offer developing countries?", *Journal of Public Health*, vol. 27, no. 2, pp. 215-221.

Middleton, P. 2004, "How allocation concealment is handled in Cochrane Reviews", *Chinese Journal of Evidence-Based Medicine*, vol. 4, no. 10, pp. 711-713.

Miles, M. B. & Huberman, A. M. 1994, *Qualitative Data Analysis* (2nd Edition). Thousand Oaks, California: Sage.

Mills, E. J., Wu, P., Gagnier, J., & Devereaux, P. J. 2005, "The quality of randomized trial reporting in leading medical journals since the revised CONSORT statement", *Contemporary Clinical Trials*, vol. 26, pp. 480-487.

Milne, R. & Oliver, S. 1996, "Evidence-based consumer health information: developing teaching in critical appraisal skills", *International Journal for Quality in Health Care*, vol. 8, no. 5, pp. 439-445.

Moher, D., Jadad, A. R., Nichol, G., Penman, M., Tugwell, P., & Walsh, S. 1995, "Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists", *Contemporary Clinical Trials*, vol. 16, no. 1, pp. 62-73.

Moher, D., Cook, D. J., Jadad, A. R., Tugwell, P., Moher, M., Jones, A., Pham, B., & Klassen, T. P. 1999, "Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses", *Health Technology Assessment*, vol. 3, no. 12.

Moher, D., Pham, B., Jones, A., Cook, D. J., Jadad, A. R., Moher, M., Tugwell, P., & Klassen, T. P. 1998, "Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses?", *Lancet*, vol. 352, no. 9128, pp. 609-613.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. 2000, "Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. QUOROM Group", *British Journal of Surgery.*, vol. 87, no. 11, pp. 1448-1454.

Moher, D., Schulz, K. F., Altman, D. G., & Lepage, L. 2001a, "The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials", *Lancet*, vol. 357, no. 9263, pp. 1191-1194.

Moher, D., Jones, A., & Lepage, L. 2001b, "Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation", *Journal of the American Medical Association*, vol. 285, no. 15, pp. 1992-1995.

Moja, L. P., Telaro, E., D'Amico, R., Moschetti, I., Coe, L., & Liberati, A. 2005, "Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study", *British Medical Journal*, vol. 330, no. 7499, p. 1053.

Morse, J. 2003, "Principles of Mixed Methods and Multimethod Research Design," in *Handbook of Mixed Methods in Social & Behavioural Research*, A. Tashakkori & C. Teddlie, eds., Sage, Thousand Oaks, California, pp. 189-208.

Mullen, P. D., Mains, D. A., & Velez, R. 1992, "A meta-analysis of controlled trials of cardiac patient education", *Patient Education & Counseling*, vol. 19, no. 2, pp. 143-162.

Mullen, P. D., Green, L. W., & Persinger, G. S. 1985, "Clinical trials of patient education for chronic conditions: a comparative meta-analysis of intervention types.", *Preventive Medicine*, vol. 14, no. 6, pp. 753-781.

Murchie, P., Hannaford, P. C., Wyke, S., Nicolson, M. C., & Campbell, N. C. 2007, "Designing an integrated follow-up programme for people treated for cutaneous malignant melanoma: a practical application of the MRC framework for the design and evaluation of complex interventions to improve health", *Family Practice.* vol. 24, no. 3, pp. 283-92.

National Institute for Health and Clinical Excellence (NICE) 2004, *Guide to the Technology Appraisal Process*. London: NICE.

National Institute for Health and Clinical Excellence 2006, *Methods for Development of NICE Public Health Guidance*. London: NICE.

National Institute for Health and Clinical Excellence (NICE). 2009. *About the Patient Public Involvement Programme.* http://www.nice.org.uk/getinvolved/patientandpublicinvolvement/patientandpublicinvolvementprogramme/patient_and_public_involvement_programme.jsp?domedia=1&mid=D4426D98-19B9-E0B5-D4C052E69B565324 [accessed 3/3/09]

National Institute for Health Research 2008, *Transforming Health Research the First Two Years - National Institute for Health Research Progress Report 2006–2008*. London: Department of Health.

Nightingale, D. & Cromby, J. 1999, *Social Constructionist Psychology : A Critical Analysis of Theory and Practice.* Buckingham: Open University Press.

Nind, M. 2006, "Conducting Systematic Review in Education", *London Review of Education*, vol. 4, no. 2, pp. 183-195.

NIHR Health Technology Assessment Programme. 2009. HTA Clinical Evaluation and Trials: an Open Call Specification Document http://www.hta.ac.uk/funding/clinicaltrials/24April09CE_specificationdocument.pdf [accessed 24/4/09]

Nurmohamed, M. T., Rosendaal, F. R., Buller, H. R., Dekker, E., Hommes, D. W., Vandenbroucke, J. P., & Briet, E. 1992, "Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis", *Lancet*, vol. 340, no. 8812, pp. 152-156.

Nutbeam, D. 2001, "Assessing the effectiveness of public health interventions, oral presentation, Evidence into Practice: Challenges and Opportunities for UK Public Health". London: The Royal College of Physicians.

Nutbeam, D. 2003, "How does evidence influence public health policy? Tackling health inequalities in England.", *Health Promotion Journal of Australia*, vol. 14, pp. 154-158.

Nutbeam, D. 1999, "Oakley's case for using randomised controlled trials is misleading", *British Medical Journal*, vol. 318, no. 7188, pp. 944-945.

Nutbeam, D. & Harris, E. 2004, *Theory in a Nutshell: A Guide to Health Promotion Theory* Sydney.

Oakley, A. 1981, "Interviewing women: a contradiction in terms (and a subsequent exchange with Joanna Malseed)," in *Doing Feminist Research*, H. Roberts, ed., Routledge & Kegan Paul, London, pp. 30-61.

Oakley, A. 1998, "Experimentation and social interventions: a forgotten but important history", *British Medical Journal*, vol. 317, no. 7167, pp. 1239-1242.

Oakley, A. 1999, "Paradigm wars: some thoughts on a personal and public trajectory", *International Journal of Social Research Methodology*, vol. 2, no. 3, pp. 247-254.

Oakley, A. 2000, *Experiments in Knowing: Gender and Method in the Social Sciences* Polity Press, Oxford.

Oakley, A. 2003, "Research Evidence, Knowledge Management and Educational Practice: early lessons from a systematic approach", *London Review of Education*, vol. 1, no. 1, pp. 21-33.

Oakley, A., Strange, V., Toroyan, T., Wiggins, M., Roberts, I., & Stephenson, J. 2003, "Using random allocation to evaluate social interventions: Three recent UK examples", *Annals of the American Academy of Political and Social Science*, vol. 589, pp. 170-189.

Oakley, A., Gough, D., Oliver, S., & Thomas, J. 2005, "The politics of evidence and methodology: lesson from the EPPI-Centre", *Evidence and Policy*, vol. 1, no. 1, pp. 5-31.

Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. 2006, "Process evaluation in randomised controlled trials of complex interventions", *British Medical Journal*, vol. 332, no. 7538, pp. 413-416.

Oakley, A., Fullerton, D., Holland, J., Arnold, S., France-Dawson, M., Kelley, P., & McGrellis, S. 1995, "Sexual health education interventions for young people: a methodological review", *British Medical Journal*, vol. 310, no. 6973, pp. 158-162.

Ogilvie, D., Egan, M., Hamilton, V., & Petticrew, M. 2005, "Systematic reviews of health effects of social interventions: 2. Best available evidence: how low should you go?", *Journal of Epidemiology & Community Health*, vol. 59, no. 10, pp. 886-892.

Oliver, S., Thomas, J., Harden, A., & Oakley, A. 2006, "Accumulating evidence to bring policy, practice and research together," in *Public Health Evidence: Tackling Health Inequalities*, A. Killoran, C. Swann, & M. Kelly, eds. Oxford: Oxford University Press. pp. 125-140.

Oliver, S. & Peersman, G. 2001, "Critical appraisal of research evidence: finding useful and reliable answers," in *Using Research for Effective Health Promotion*, S. Oliver & G. Peersman, eds. Buckingham: Open University Press, pp. 82-95.

Oliver, S., Bagnall, A., Thomas, J., Shepherd, J., Sowden, A., White, I., Dines, J., Rees, R., Colquitt, J., Oliver, K., & Garret, Z. 2008, *RCTs for policy interventions?: A Review of Reviews and Meta-Regression. Final report to the National Coordinating Centre for Research*

Methodology (NCCRM), Institute of Education / University of Southampton / University of York.

Oliver, S., Clarke-Jones, L., Rees, R., Milne, R., Buchanan, P., Gabbay, J., Gyte, G., Oakley, A., & Stein, K. 2004, "Involving consumers in research and development agenda setting for the NHS: developing an evidence-based approach.", *Health Technology Assessment*, vol. 8, no. 15.

Oppenheim, A. N. 1992, *Questionnaire Design, Interviewing and Attitude Measurement*. London: Pinter.

Øvretveit, J. 1998, *Evaluating Health Interventions: An Introduction to Evaluation of Health Treatments, Services, Policies and Organisational Interventions* Open University Press.

Paolucci-El Dib, R., Atallah, A., & Andriolo, R. B. 2005, "Mapping the Cochrane evidence for decision-making in health care [abstract]", *XIII Cochrane Colloquium. 22nd to 26th October. Melbourne., Australia*.

Parkes, J., Hyde, C., Deeks, J., & Milne, R. 2001, "Teaching critical appraisal skills in health care settings.", *Cochrane Database of Systematic Reviews* no. 3, p. CD001270.

Patton, M. 1990, *Qualitative Research and Evaluation Methods*. California: Sage.

Pawson, R. 2006a, *Evidence-Based Policy : A Realist Perspective.* London: Sage.

Pawson, R. 2006b, "Simple principles for the evaluation of complex programmes," in *Public Health Evidence: Tackling Health Inequalities*, A. Killoran, C. Swann, & M. Kelly, eds. Oxford: Oxford University Press, pp. 223-238.

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. 2005, "Realist review--a new method of systematic review designed for complex policy interventions", *Journal of Health Services. Research and Policy*, vol. 10 Suppl 1, pp. 21-34.

Peersman, G., Harden, A., & Oliver, S. 1999, *Effectiveness Reviews in Health Promotion*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Peersman, G., Oliver, S., & Oakley, A. 1997, *EPPI-Centre Review Guidelines*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Petrosino, A., Turpin-Petrosino, C., & Buehler, J. 2002, *"Scared Straight" and Other Juvenile Awareness Programmes for Preventing Juvenile Delinquency*. Philadelphia: The Campbell Collaboration.

Petticrew, M. & Roberts, H. 2006, *Systematic Reviews in the Social Sciences: A Practical Guide.* Oxford: Blackwell.

Petticrew, M., Whitehead, M., Bambra, C., Egan, M., Graham, H., Macintyre, S., & McDermott, E. 2006, "The Centre for Evidence-Based Public Health Policy: part of the ESRC Evidence Network," in *Public Health Evidence: Tackling Health Inequalities*, A. Killoran, C. Swann, & M. Kelly, eds. Oxford: Oxford University Press, pp. 141-154.

Petticrew, M., Whitehead, M., Macintyre, S. J., Graham, H., & Egan, M. 2004, "Evidence for public health policy on inequalities: 1: the reality according to policymakers", *Journal of Epidemiology & Community Health*, vol. 58, no. 10, pp. 811-816.

Petticrew, M. 2003, "Why certain systematic reviews reach uncertain conclusions.", *British Medical Journal.*, vol. 326, no. 7392, pp. 756-758.

Pettifor, A. E., MacPhail, C., Bertozzi, S., & Rees, H. V. 2007, "Challenge of evaluating a national HIV prevention programme: the case of loveLife, South Africa", *Sexually Transmitted Infections*, vol. 83, no. suppl 1, p. i70-i74.

Piggott, M., McGee, H., & Feuer, D. 2004, "Has CONSORT improved the reporting of randomized controlled trials in the palliative care literature? A systematic review.", *Palliative Medicine*, vol. 18, pp. 32-38.

Pildal, J., Hrobjartsson, A., Jorgensen, K. J., Hilden, J., Altman, D. G., & Gotzsche, P. C. 2007, "Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials", *International Journal of Epidemiology*, vol. 36, no. 4, pp. 847-857.

Plint, A. C., Moher, D., Morrison, A., Schulz, K., Altman, D. G., Hill, C., & Gaboury, I. 2006, "Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review", *Medical Journal of Australia*, vol. 185, pp. 263-267.

Polanyi, M. 1966, *The Tacit Dimension.* New York: Anchor Day Books.

Powell, C., Kavanagh, J., Brunton, G., Rees, R., & Waters, E. 2005, "Increasing specificity in searches of PubMed for health promotion and public health studies: a prospective study [abstract]", *XIII Cochrane Colloquium. 22nd to 26th October. Melbourne., Australia*.

Powell, C., Wedner, S., & Richardson, S. 2004, "Screening for correctable visual acuity deficits in school-age children and adolescents.", *Cochrane Database of Systematic Reviews* no. 1, p. CD005023.

Prochaska, J. O., Velicer, W. F., Rossi, J. S., Goldstein, M. G., Marcus, B. H., Rakowski, W., Fiore, C., Harlow, L. L., Redding, C. A., Rosenbloom, D., & et, a. 1994, "Stages of change and decisional balance for 12 problem behaviors", *Health Psychology.*, vol. 13, no. 1, pp. 39-46.

Protheroe, J., Bower, P., & Chew-Graham, C. 2007, "The use of mixed methodology in evaluating complex interventions: identifying patient factors that moderate the effects of a decision aid", *Family Practice.,* vol. 24, no. 6, pp. 594-600.

Rees, R., Kavanagh, J., Burchett, H., Shepherd, J., Brunton, G., Harden, A., Thomas, J., Oliver, S., & Oakley, A. 2004a, *HIV Health Promotion and Men Who Have Sex With Men (MSM): A Systematic Review of Research Relevant to the Development and Implementation of Effective and Appropriate Interventions*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Rees, R., Oliver, S., Harden, A., Shepherd, J., Kavanagh, J., Burchett, H., Brunton, G., Thomas, H., & Oakley, A. 2004b, "Use of an advisory group to ensure relevance: reflections on participation of stakeholders in a review of sexual health promotion for men who have sex with men (MSM)" *12th Cochrane Colloquium*, Ottawa, Canada, October 2nd - 6th.

Rees, R., Kavanagh, J., Harden, A., Shepherd, J., Brunton, G., Oliver, S., & Oakley, A. 2006, "Young people and physical activity: a systematic review matching their views to effective interventions", *Health Education Research*, vol. 21, no. 6, pp. 806-825.

Richards, L. 1999, *Using NVivo in Qualitative Research* Sage, Thousand Oaks, California.

Robson, C. 2002, *Real World Research: a Resource for Social Scientists and Practitioner-Researchers* (2nd Edition). Oxford: Blackwell.

Rossi, P., Freeman, H., & Lipsey, M. 1999, *Evaluation : A Systematic Approach*, 2nd edition. California, Thousand Oaks: Sage Publications.

Rotheram-Borus, M. J., Cantwell, S., & Newman, P. A. 2000, "HIV prevention programs with heterosexuals", *AIDS*, vol. 14, no. 2 Supplement, p. S59-S67.

Rothman, A. J. 2004, ""Is there nothing more practical than a good theory?": Why innovations and advances in health behavior change will arise if interventions are used to test and refine theory", *International Journal of Behavioral Nutrition and Physical Activity*, vol. 1, no. 11.

Royle, J. & Oliver, S. 2004, "Consumer involvement in the health technology assessment program", *International Journal of Technology Assessment in Health Care*, vol. 20, no. 4, pp. 493-497.

Rubin, H. & Rubin, I. 2005, *Qualitative Interviewing: The Art of Hearing Data* (2nd Edition). California, Thousand Oaks: Sage Publications.

Ruiz-Canela, M., Irala-Estevez, J., Martinez-Gonzalez, M. A., Gomez-Gracia, E., & Fernandez-Crehuet, J. 2001, "Methodological quality and reporting of ethical requirements in clinical trials.", *Journal of Medical Ethics.*, vol. 27, no. 3, pp. 172-176.

Rush, B., Shiell, A., & Hawe, P. 2004, "A census of economic evaluations in health promotion.", *Health Education Research*, vol. 19, no. 6, pp. 707-719.

Rychetnik, L., Frommer, M., Hawe, P., & Shiell, A. 2002, "Criteria for evaluating evidence on public health interventions", *Journal of Epidemiology & Community Health.*, vol. 56, no. 2, pp. 119-127.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. 1996, "Evidence based medicine: what it is and what it isn't", *British Medical Journal*, vol. 312, no. 7023, pp. 71-72.

Sanders, J. 1999, "An evaluation of "The effectiveness of comprehensive, case management interventions: evidence from the national evaluation of the comprehensive child development program"", *American Journal of Evaluation*, vol. 20, no. 3, pp. 577-582.

Sarnoff, R. & Rundall, T. 1998, "Meta-analysis of effectiveness of interventions to increase influenza immunization rates among high-risk population groups", *Medical Care Research and Review*, vol. 55, no. 4, pp. 432-456.

Schulz, K. F., Chalmers, I., Grimes, D. A., & Altman, D. G. 1994, "Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals", *Journal of the American Medical Association*, vol. 272, no. 2, pp. 125-128.

Schulz, K. F., Chalmers, I., Hayes, R. J., & Altman, D. G. 1995, "Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials", *Journal of the American Medical Association*, vol. 273, no. 5, pp. 408-412.

Schulz, K. F. 1995, "Subverting randomization in controlled trials", *Journal of the American Medical Association*, vol. 274, no. 18, pp. 1456-1458.

Schulz, K. F. & Grimes, D. A. 2002, "Generation of allocation sequences in randomised trials: chance, not choice", *Lancet*, vol. 359, no. 9305, pp. 515-519.

Seidel, J. & Clarke, J. 1984, "The ETHNOGRAPH: a computer programe for the analysis of qualitative data.", *Qualitative Sociology*, vol. 7, no. 1-2, pp. 110-25.

Senn, S. 1994, "Testing for baseline balance in clinical trials", *Statistics in Medicine*, vol. 13, no. 17, pp. 1715-1726.

Shacklock, G. & Smyth, J. 1998, *Being Reflexive in Critical Educational and Social Research.* London: Falmer Press.

Shadish, W. R., Navarro, A. M., Matt, G. E., & Phillips, G. 2000, "The effects of psychological therapies under clinically representative conditions: a meta-analysis", *Psychological Bulletin*, vol. 126, no. 4, pp. 512-529.

Sheldon, T. A., Cullum, N., Dawson, D., Lankshear, A., Lowson, K., Watt, I., West, P., Wright, D., & Wright, J. 2004, "What's the evidence that NICE guidance has been implemented? Results from a national evaluation using time series analysis, audit of patients' notes, and interviews.", *British Medical Journal*, vol. 329, no. 7473.

Shepherd, J. 1998, *An investigation into effective peer-led HIV prevention and sexual health promotion with young gay and bisexual men. Thesis (M. Phil.).* Southampton: School of Education, University of Southampton.

Shepherd, J. 2009. “Assessing quality in systematic reviews of the effectiveness of health promotion and public health: Areas of consensus and dissension”. Abstract for oral presentation submitted to the *17th Cochrane Colloquium*, Singapore, 11th to 14th October 2009.

Shepherd, J., Kavanagh, J., Picot, J., Cooper, K., Harden, A., Barnett-Page, E., Jones, J., Clegg, A., Hartwell, D., Frampton, G., & Price, A. 2009, "The effectiveness and cost-effectiveness of behavioural interventions for the prevention of sexually transmitted infections in young people

aged 13 to 19: a systematic review and economic evaluation", *Health Technology Assessment*, vol. In Press.

Shepherd, J., Rogers, G., Anderson, R., Main, C., Thompson-Coon, J., Hartwell, D., Liu, Z., Loveman, E., Green, C., Pitt, M., Stein, K., Harris, P., Frampton, G., Smith, M., Takeda, P., Price, A., Welch, K., & Somerville, M. 2008b, "Inhaled corticosteroids and Long Acting Beta-2 Agonists for the treatment of chronic asthma in adults and children aged 12 years and over: Systematic review and economic analysis", *Health Technology Assessment*, 2008 vol. 12, no. 19.

Shepherd, J., Briggs, J., Payne, E., Packer, C., Kerridge, L., & Ashton-Key 2007a, "Setting the future policy agenda for health technology assessment: a specialty mapping approach", *International Journal of Technology Assessment in Health Care*, vol. 23, no. 4.

Shepherd, J., Jones, J., Hartwell, D., Davidson, P., Price, A., & Waugh, N. 2007b, "Interferon alfa (pegylated and non-pegylated) and ribavirin for the treatment of mild chronic hepatitis c - a systematic review and economic evaluation", *Health Technology Assessment*, vol. 11, no. 11.

Shepherd, J., Harden, A., Rees, R., Brunton, G., Garcia, J., Oliver, S., & Oakley, A. 2006a, "Young people and healthy eating: a systematic review of research on barriers and facilitators", *Health Education Research: Theory and Practice*, vol. 21, no. 2, pp. 239-257.

Shepherd, J., Jones, J., Takeda, A., Davidson, P., & Price, A. 2006b, "Adefovir dipivoxil and pegylated interferon alfa-2a for the treatment of chronic hepatitis B - a systematic review and economic evaluation", *Health Technology Assessment*, vol. 10, no. 28.

Shepherd, J., Brodin, H., Cave, C., Waugh, N., Price, A., & Gabbay, J. 2004, "Pegylated interferon alpha 2a and 2b in combination with ribavirin in the treatment of chronic hepatitis C : a systematic review", *Health Technology Assessment*, vol. 8, no. 39.

Shepherd, J., White, I., Rees, R., Thomas, J., Brunton, G., Harden, A., Kavanagh, J., Sutcliffe, K., Oliver, S., & Oakley, A. 2003, "A systematic comparison of different sets of quality assessment criteria in systematic reviews of effectiveness in health promotion. *XI Cochrane Colloquium*, Barcelona, 26th – 31st October.

Shepherd, J. & Harden, A. 2003, "The value of systematic reviews of the effectiveness of sexual health interventions," in *Effective sexual health interventions: issues in experimental evaluation*, J. Imire, J. Stephenson, & C. Bonell, eds. Oxford: Oxford University Press.

Shepherd, J., Harden, A., Rees, R., Brunton, G., Oliver, S., & Oakley, A. 2001 "Synthesising evidence from different study types : systematic reviews on the barriers and facilitators to the health of young people", *9th Cochrane Colloquium, 9th – 13th October. Lyon, France.*

Shepherd, J., Waugh, N., & Hewitson, P. 2000, "Combination therapy (interferon alfa and ribavirin) in the treatment of chronic hepatitis C: a rapid and systematic review", *Health Technology Assessment.*, vol. 4, no. 33.

Shepherd, J., W. R., Peersman, G., & Napuli, I. "Cervical cancer and sexual lifestyle : a systematic review of health education interventions - *VII Cochrane Colloquium*", 1999a, Rome.

Shepherd, J., Turner, G., & Weare, K. 1999b, "A new method of peer-led HIV prevention with gay and bisexual men," in *AIDS: Families, Culture and Community*, P. Aggleton, P. Davies, & G. Hart, eds. London: Taylor and Francis, pp. 163-183.

Shepherd, J., Turner, G., & Weare, K. 1997a, *Sexual Health Promotion with Young Gay and Bisexual Men: A New Method of Working*. Southampton: School of Education, University of Southampton.

Shepherd, J., Weare, K., & Turner, G. 1997b, "Peer-led sexual health promotion with young gay and bisexual men results of The HAPEER Project", *Health Education*, vol. 6, pp. 204-212.

Shults, R. A., Elder, R. W., Sleet, D. A., Nichols, J. L., Alao, M. O., Carande-Kulis, V. G., Zaza, S., Sosin, D. M., Thompson, R. S., & Task Force on Community Preventive Services 2001, "Reviews of evidence regarding interventions to reduce alcohol-impaired driving. ", *American Journal of Preventive Medicine*, vol. 21, no. 4:Suppl, p. Suppl-88.

Shufflebeam, D. 2000, "The methodology of meta evaluation as reflected in metaevaluations by the Western Michigan University Evaluation Centre", *Journal of Personnel Evaluation in Education*, vol. 14, pp. 95-125.

Silva, M. 2002, "The effectiveness of school-based sex education programs in the promotion of abstinent behavior: a meta-analysis", *Health Education Research.*, vol. 17, no. 4, pp. 471-481.

Sindhu, F., Carpenter, L., & Seers, K. 1997, "Development of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique", *Journal of Advanced Nursing*, vol. 25, no. 6, pp. 1262-1268.

Slavin, R. E. 1990, "Ability grouping and student achievement in secondary schools: A best-evidence synthesis", *Review of Educational Research*, vol. 60, no. 471, p. 499.

Slavin, R. E. 1986, "Best-evidence synthesis: an alternative to meta-analytic and traditional reviews", *Educational Researcher*, vol. 15, no. 9, pp. 5-11.

Slavin, R. E. 1995, "Best evidence synthesis: an intelligent alternative to meta-analysis", *Journal of Clinical Epidemiology.*, vol. 48, no. 1, pp. 9-18.

Sowden, A. & Glanville, J. 2006, "The Centre for Reviews and Dissemination," in *Public Health Evidence: Tackling Health Inequalities*, A. Killoran, C. Swann, & M. Kelly, eds. Oxford: Oxford University Press, pp. 95-114.

Speller, V., Learmonth, A., & Harrison, D. 1997, "The search for evidence of effective health promotion", *British Medical Journal*, vol. 315, no. 7104, pp. 361-363.

Stephenson, J. M., Oakley, A., Charleston, S., Brodala, A., Fenton, K., Petruckevitch, A., & Johnson, A. M. 1998, "Behavioural intervention trials for HIV/STD prevention in schools: are they feasible?", *Sexually Transmitted Infections*, vol. 74, no. 6, pp. 405-408.

Stephenson, J. & Imrie, J. 1998, "Why do we need randomised controlled trials to assess behavioural interventions?", *British Medical Journal*, vol. 316, no. 7131, pp. 611-613.

Stephenson, J. M., Strange, V., Forrest, S., Oakley, A., Copas, A., Allen, E., Babiker, A., Black, S., Ali, M., Monteiro, H., Johnson, A. M., & RIPPLE study team 2004, "Pupil-led sex education in England (RIPPLE study): cluster-randomised intervention trial.", *Lancet.364(9431):338-46,* p. -30.

Stewart, A. 2002, *Basic Statistics and Epidemiology.* Oxford: Radcliffe Medical Press.

Stout J & Rivara F 1989, "Schools sex education: does it work?", *Pediatrics*, vol. 83, no. 3, pp. 375-379.

Strauss, A. L. & Corbin, J. 1990, *Basics of Qualitative Research: Grounded Theory Procedures and Techniques* Sage. California: Newbury Park.

Swann, C., Bowe, K., McCormick, G., & Kosmin, M. 2003, *Teenage Pregnancy and Parenthood: a Review of Reviews (Evidence Briefing)*. London: Health Development Agency.

Tallon, D., Jüni, P., & Egger, M. "Assessment of the quality of controlled trials in meta-analyses published in leading journals", *16th Annual Meeting of the International Society of Technology Assessment in Health Care*", 18th -2 1st June. The Hague, Netherlands.

Tashakkori, A. & Teddlie, C. 2003, *Handbook of Mixed Methods in Social & Behavioural Research* Sage. California: Thousand Oaks.

Teenage Pregnancy Independent Advisory Group 2008, *Annual Report 2007/8*. London: Department for Children, Schools and Families (DCSF).

Thomas, J. EPPI-Reviewer© 2.0 (Web edition). EPPI-Centre Software. 2002. London: Social Science Research Unit, Institute of Education, University of London.

Thompson, S. G. & Pocock, S. J. 1991, "Can meta-analyses be trusted?", *Lancet*, vol. 338, no. 8775, pp. 1127-1130.

Thomson, H., Hoskins, R., Petticrew, M., Ogilvie, D., Craig, N., Quinn, T., & Lindsay, G. 2004, "Evaluating the health effects of social interventions", *British Medical Journal.*, vol. 328, no. 7434, pp. 282-285.

Thomson, H., Jepson, R., Hurley, F., & Douglas, M. 2008, "Assessing the unintended health impacts of road transport policies and interventions: translating research evidence for use in policy and practice.", *BMC Public Health*, vol. 8, p. 339.

Tierney, J. & Stewart, L. 2005, "Investigating patient exclusion bias in meta-analysis", *International Journal of Epidemiology*, vol. 34, no. 79, p. 87.

Tilford, S. 2000, "Evidence-based health promotion (Editorial)", *Health Education Research.*, vol. 15, no. 6, pp. 659-663.

Tingle, L. R. 2003, "Short title A meta-evaluation of 11 school-based smoking prevention program", *Journal of School Health*, vol. 73, no. 2, pp. 64-67.

Tones, K. 2000, "Evaluating health promotion: a tale of three errors", *Patient Education and Counselling*, vol. 39, no. 2-3, pp. 227-236.

Tones, K. & Tilford, S. 2001, *Health Promotion: effectiveness, efficiency and equity (3rd Edition)*. Cheltenham: Nelson Thornes.

Torgerson, C. 2003, *Systematic reviews.* London: Continuum.

Torgerson, D. J. 2001, "Contamination in trials: is cluster randomisation the answer?", *British Medical Journal*, vol. 322, no. 7282, pp. 355-357.

Turner JA, Schroth WS, & Fordyce WE 1996, "Educational and behavioural interventions for back pain in primary care", *Spine*, vol. 21, pp. 2851-2859.

Turner, G. & Shepherd, J. 1999, "A method in search of a theory: peer education and health promotion, *Health Education Research.*, vol. 14, no. 2, pp. 235-247.

van Driel, W. G. & Keijsers, J. F. 1997, "An instrument for reviewing the effectiveness of health education and health promotion", *Patient Education & Counseling.*, vol. 30, no. 1, pp. 7-17.

van Tulder, M. W., Ostelo, R., Vlaeyen, J. W., Linton, S. J., Morley, S. J., & Assendelft, W. J. 2001, "Behavioral treatment for chronic low back pain: a systematic review within the framework of the Cochrane Back Review Group.", *Spine*, vol. 26, no. 3, pp. 270-281.

Vartiainen, E., Tossavainen, K., Viri, L., Niskanen, E., & Puska, P. 1991, "The North Karelia Youth Programmes," in *Hyperlipidemia in Childhood and the Development of Atherosclerosis*, C. Williams & E. Wynder, eds. New York: Annals of the New York Academy of Science.

Verhagen, A. P., de Vet, H. C., de Bie, R. A., Kessels, A. G., Boers, M., Bouter, L. M., & Knipschild, P. G. 1998, "The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus", *Journal of Clinical Epidemiology.*, vol. 51, no. 12, pp. 1235-1241.

Wallace, A., Croucher, K., Bevan, M., Jackson, K., O'Malley, L., & Quilgars, D. 2006, "Evidence for policy making: Some reflections on the application of systematic reviews to housing research", *Housing Studies*, vol. 21, no. 2, pp. 297-314.

Walley, T. & Thakker, R. V. 2008, "Developments for funding clinical research in the UK", *Lancet*, vol. 372, no. 9638, pp. 518-519.

Wanless, D. 2004, *Securing Good Health for the Whole Population*. London: Her Majesty's Stationery Office (HMSO).

Waters, E., Doyle, J., Jackson, N., Howes, F., Brunton, G., Oakley, A., & Cochrane, C. 2006, "Evaluating the effectiveness of public health interventions: the role and activities of the Cochrane Collaboration", *Journal of Epidemiology & Community Health*, vol. 60, no. 4, pp. 285-289.

Waugh, N., Royle, P., Robertson, L., Shepherd, J., & Loveman, E. "Should there be Cochrane reviews of disease aetiology? A discussion poster". *XIII Cochrane Colloquium, 22nd to 26th October. Melbourne., Australia.*

Weber, P. 1990, *Basic Content Analysis (2nd Edition).* Thousand Oaks, California: Sage.

Weightman, A., Ellis, S., Cullum, A., Sander, L., & Turley, R. 2005, *Grading Evidence and Recommendations for Public Health Interventions: Developing and Piloting a Framework*, Health Development Agency, London.

Weingarten, M., Paul, M., & Leibovici, L. 2004, "Assessing ethics of trials in systematic reviews", *British Medical Journal*, vol. 328, no. 7446, pp. 1013-1014.

Weitzman, E. A. 2008, "Software and qualitative research," in *Handbook of Qualitative Research (2nd Edition)*, N. K. Denzin & Y. S. Lincoln, eds. Thousand Oaks, California: Sage, pp. 803-820.

West, S., King, V., Carey, T., Lohr, K., McKoy, N., Sutton, S., & Lux, L. 2002, *Systems to Rate the Strength Of Scientific Evidence*. Rockville, MD: Agency for Healthcare Research and Quality, Evidence Report/Technology Assessment 47.

White, D. 2001, "Evaluating evidence and making judgements of study quality: loss of evidence and risks to policy and practice decisions", *Critical Public Health*, vol. 11, no. 1, pp. 3-17.

White, I. R. & Thomas, J. 2005, "Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis", *Clinical Trials*, vol. 2, no. 2, pp. 141-151.

Whitehead, M. 1996, "Evidence based health education (editorial)", *Health Education Journal*, vol. 55, no. 1, pp. 1-2.

WHO European Working Group on Health Promotion Evaluation 1998, *Health Promotion Evaluation: Recommendation to Policymakers*. Copenhagen: WHO Regional Office for Europe.

Wight, D., Abraham, C., & Scott, S. 1998, "Towards a psycho-social theoretical framework for sexual health promotion, *Health Education Research*, vol. 13, no. 3, pp. 317-330.

Wiles, R. & Bardsley, N. 2008, *ESRC National Centre for Research Methods Evaluating the impact of NCRM Training and Capacity Building Activities*. Southampton: University of Southampton.

Wilson, S. J. & Lipsey, M. W. 2000, "Wilderness Challenge Programs for Delinquent Youth: A Meta-Analysis of Outcome Evaluations", *Evaluation and Program Planning*, vol. 23, no. 1, pp. 1-12.

Wolitski RJ, MacGowan RJ, Higgins DL, & Jorgensen CM 1997, "The effects of HIV counseling and testing on risk-related practices and help-seeking behavior", *AIDS Education and Prevention*, vol. 9, no. supplement B, pp. 52-67.

Wood, L., Egger, M., Gluud, L. L., Schulz, K. F., Juni, P., Altman, D. G., Gluud, C., Martin, R. M., Wood, A. J., & Sterne, J. A. 2008, "Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study", *British Medical Journal*, vol. 336, no. 7644, pp. 601-605.

World Health Organisation 1986, *Ottawa Charter for Health Promotion*. Geneva: World Health Organisation.

World Health Organisation 1978, *Report on the international conference on Primary Health Care, Alma Ata, 6-12 September.* Geneva: World Health Organisation.

World Health Organisation 1997, *The Jakarta Declaration on Health Promotion into the 21st Century*. Geneva: World Health Organisation..

Zaza, S., Wright-De Aguero, L. K., Briss, P. A., Truman, B. I., Hopkins, D. P., Hennessy, M. H., Sosin, D. M., Anderson, L., Carande-Kulis, V. G., Teutsch, S. M., & Pappaioanou, M. 2000, "Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task Force on Community Preventive Services", *American Journal of Preventive Medicine*, vol. 18, no. 1:Suppl, p. Suppl-74.

Zhu, S. H., Anderson, C. M., Tedeschi, G. J., Rosbrook, B., Johnson, C. E., Byrd, M., & Gutierrez-Terrell, E. 2002, "Evidence of real-world effectiveness of a telephone quitline for smokers.", *New England Journal of Medicine*, vol. 347, no. 14, pp. 1087-1093.