BEME GUIDE

Conducting a best evidence systematic review. Part 1: From idea to data coding. BEME Guide No 13

Marilyn Hammick, BEME, UK
Timothy Dornan, University of Manchester, UK
Yvonne Steinert, McGill University, Montreal, Canada

Notes on contributors:

Marilyn Hammick is a Research and Education Consultant, Visiting Professor at Birmingham City & Anglia Ruskin Universities and Consultant to Best Evidence Medical Education. She is a member of the WHO Study Group on Interprofessional Education and Collaborative Care, past Chair of the UK Centre for Interprofessional Education and Learning & Teaching Consultant for Research and Evaluation in Health Sciences & Practice at the UK Higher Education Academy. Other roles include associate editor, The Journal of Interprofessional Care and editorial board member, Medical Teacher.

Tim Dornan is Professor of Medicine and Clinical Education at University of Manchester, UK, and Honorary Consultant Physician, Salford Royal Hospital. He was Convener of the BEME Review Group on 'How can experience in clinical and community settings contribute to early medical education?' and currently leads the BEME Review Group conducting 'A review of the evidence linking conditions, processes and outcomes of clinical workplace learning.'

Yvonne Steinert, a clinical psychologist, is Professor of Family Medicine, Associate Dean for Faculty Development and the Director of the Centre for Medical Education at McGill University, Montreal, Canada. She led the BEME Review Group that focused on 'A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education'.

Practice points

- A clear review question lies at the heart of systematic review research.
- A pilot review ensures that the review outcomes are robust and acceptable to the community of practice relevant to the topic under review.
- Systematic review quality is related to the quality of the primary evaluations; review research includes making a judgement on the quality of these using an appropriate tool.
- A cohesive, committed and appropriately experienced review group ensures that conducting a systematic review is feasible and satisfactory.
- Good practice in systematic review work focuses on undertaking tasks systematically, consistently and reporting this work in ways that are transparent

Introduction

Evidence informed practice is increasingly the norm for many professionals, emerging from the seminal discourse of evidence based medicine, through the work of (mainly) the Cochrane Collaboration (Hammick 2005). Evidence based practice is seen as the way to link knowledge from either primary research or systematic reviews and the logical and reliable application of that knowledge in professional practice. In keeping with this movement, systematic reviews the their aim of providing evidence about the effectiveness of an education intervention, are now contributing to knowledge about health care sciences education and providing a valuable resource for education practitioners and policy makers.

Since its inception in 1999, The Best Evidence Medical Education Collaboration (BEME) has played a major international role in this work by supporting systematic review work and the dissemination of evidence to education practitioners and policy makers (Harden et al 1999; Harden & Lilley 2000). At that time, Hart and Harden (2000) observed that there should be a move towards the ethos and practice of best evidence medical education (BEME). That move is well underway and so it is timely to produce specific guidance to enhance the quality of future BEME systematic reviews.

High level and well targeted engagement with selected primary studies can provide evidence that, clearly and succinctly reported, enables clinical and classroom based education practitioners to develop and change their ways of delivering learning. Equally important is providing evidence to meet the needs of those who make policy decisions regarding professional education. This includes assisting regulatory and funding bodies in decisions about curriculum content and design, quality assurance for the teaching faculty when resources are limited and priorities contentious.

Undertaking systematic review research is also invaluable to the individuals involved. Ideas for review topics often arise from a local, even personal search for evidence faced with making changes to education practices. The following quotes from BEME reviewers represent the views of many others:

- *We could find no clearly articulated rationale, and no summary of the empirical evidence. It seemed extraordinary that such major curriculum change should be made without even knowing how to evaluate it.*

- *It was an opportunity to develop our careers and show our universities evidence of our academic worth.*

- *Doing the review was about the intrinsic rewards of scholarship.*

- *We enjoyed reading scholarly publications and our involvement in the review process has been invaluable in the development of individual critical appraisal skills and writing skills.*

- *Its satisfying that our review work has led to clear improvements in the rigour of evaluations and increased interest in doing research that can contribute to the evidence future base.*

- *I enjoyed the buzz of being part of an international team, even by email.*

- *We were all involved in medical education so the question was personally important.*

Aim and content

The aims of this Guide (No. 13) and its companion Guide No. 14 (in preparation) are to set out ways of undertaking essential aspects of a systematic review of an educational topic and to discuss the conceptual reasons for these. Here you will find discussion about the systematic review process from the point of the initial idea of a topic to when all data from the selected primary studies has been coded. Guide No. 14 looks at ways of analysing and synthesising that data, good practice in making judgements from that work and in reporting secondary reviews. Both Guides will also be of value to those wishing to conduct a more modest literature review.

We have drawn extensively on the wisdom and experience of those who have undertaken systematic reviews of professional education, including the first published

BEME reviews. The use of material from completed reviews serves to illustrate the practical application of the review process discussed. In this way the guide provides practical help to new review groups and seeks to contribute to the debate about ways of obtaining evidence (and what sort of evidence) to inform policy and practice in education.

In the following pages, you will find sections about

- Forming a review group
- Clarifying the review question
- Conceptual frameworks for BEME systematic reviews
- The review protocol
- The pilot review
- Managing citations, abstracts and primary studies
- Literature and data management
- Appraising and rating the quality of primary studies
- Developing a data coding sheet
- Classifying and measuring effectiveness
- Quality enhancement procedures
- Training and development for systematic review research.

In Guide No. 14 the processes of analysing and synthesising, drawing conclusions from that work and reporting the review work in a logical, succinct and readable manner are considered.

The nature of a BEME review group

Undertaking a systematic review means a commitment to considerable work. For some review groups, the time taken to complete a review is years rather than months; BEME also asks that groups update their review at appropriate intervals. Dissemination with the aim of getting evidence for effective practice is an important part of the BEME process so readable reports and summaries that take into account the audience for the evidence are vital. Past and current BEME reviewers agree that this work is intellectually demanding and worthwhile.

Forming a cohesive group and organising the work of the review is key to making the task feasible and satisfying for everyone involved. It's important to select individuals with a track record of working on team projects and who are able to complete tasks in a timely manner. In this section, we will focus on the process of establishing a review group, with experience and examples from past groups.

Forming a review group

Each BEME review group comes into being differently. For some, it is the natural alliance of colleagues who have previously worked together; other groups begin with a leader who then gathers together like-minded and willing colleagues. At the same time, some groups have grown out of an expression of interest via the BEME Listserv and members work together for the first time during the review. It is not possible for only one person to conduct a BEME review, and we would recommend at least 4 people in the working team. Many review groups have 6-8 members which shares the workload without extending the team too much. With an even number of reviewers working in pairs to select abstracts and papers is simple to arrange.

In addition to the working team, many review groups have a wider advisory group to consult with at key points. For groups that wish to remain primarily national, inviting international colleagues to be part of the wider group is a good idea. As you can read in the following paragraph there are advantages and disadvantages in international composition of the main review group.

The composition of the Faculty Development review group was extremely international, with representatives from 6 countries (Canada, United States, Australia, Argentina, United Kingdom and Holland). Only two members came from Canada and there was no core group that worked together in the same location. Although the international composition of the group provided a rich and diverse perspective, it was difficult to coordinate group efforts and arrange group meetings. Moreover, although a subset of the group was able to meet on several occasions during other international meetings, the whole group did not meet face-to-face.

International members of the Faculty Development review translated non-English language primary studies and their review abstract is now available in Spanish and French. However, note this group's comment on the challenges of

communication with international membership. Although much can be achieved with email and phone meetings, many groups also comment on the value of at least one or two face to face working meetings, and some have bid for funds to cover the costs of travel for this. Review groups are left to make their own choices about matters such as these, but are advised to weigh up the pros and cons of their group composition early in the life of their review.

For most of the published BEME reviews members of the review group received no payment and completed the work as part of their work-related scholarly activities. For example, the Early Clinical Experience review group did not pay review group members. However, they paid a medical student who was trained as a Cochrane reviewer to do some of the electronic searching late in the process, for the translation of three articles – about 700 Euro – which cost much more than the whole of the rest of the study put together (in direct costs, at least) and 350 Euro for help to set up an Access database and enter data. This is typical of the funding of the majority of BEME review groups but there are exceptions.

The review entitled 'A systematic review of the literature on assessment, feedback and physicians' clinical performance' was conceived and supported by the American Board of Internal Medicine Foundation: the funders then selected one medical education centre to undertake the review (Veloski et al. (2006). The group conducting a review of final examinations for veterinary undergraduates at Edinburgh University, UK was successful in their bid for funding for their review. This enabled allocation of a 0.2 WTE research assistant for the duration of the review, training workshops and travel for an initial meeting of the wider constituency related to the review topic and some essential face to face meetings. It is useful for review groups to meet face to face, but may be impractical: some groups have successfully completed their review using telephone and electronic communication only. How the review group will communicate and whether a minimum number of face to face meetings are thought necessary need to be considered when forming the group (see comments on group size and composition above).

Reviewing papers and contributing to the analysis and synthesis of the secondary data is time consuming: estimates of this will of course vary. Experience to date is that reviews can take as long as two years with *part-time* reviewers and be completed in six months if a full time research assistant is part of the team. The time taken for the review to be completed is also dependant on the scope of the research

question and subsequent number of papers that qualify for review. Narrowing the review question can achieve a more manageable set of papers and thus workload.

It is worthwhile remembering that systematic reviews are similar to other large projects and use of a project planning tools, e.g. Gantt charts, that permit plotting of timescales, responsibilities etc., can help to put into perspective what will be needed for a particular research questions and how realistic these are.

Working in large and small groups

A systematic review is primarily a group activity (Reeves et al., 2002). Whatever the size and composition of your review group someone needs to lead the review work and to convene meetings. This can be the same person throughout which is efficient but arduous; alternatively some groups decide that different members can lead at different times during the work of the review. One comment from a BEME review group leader, whose views were elicited for this Guide, wrote that 'it was extremely arduous to coordinate the entire process but it was probably also fairly efficient; a small amount of paid time from a highly skilled administrator with clerical support would have been a great advantage.'

Use of research assistants/associates in review work

As we indicate above, assistance with specific tasks can help to move the review process considerably. For example, the Faculty Development review group hired a part-time research assistant (with funds solicited by the review group chair from local pharmaceutical representatives) to perform the following tasks: retrieval of articles; entry of all references into centralized database; entry of coding results into central database; follow-up with the review group members. This help was invaluable in moving the work forward. Another group used assistance to file reprints and keep an audit trail, receive coding forms and forward them for second coding, and enter data to a database.

Working with colleagues in information science and statistics

A successful systematic review is dependent on a careful literature search and expertise in this area is essential to locating appropriate and relevant primary studies. Experience in compiling and implementing a search strategy may reside in a

member of the group; an alternative is to ask a colleague in the information sciences to participate. It is essential invite your information science colleagues to join the review group early in the review process and to work with them to compile and refine the search strategy.

Similar expertise in statistics is valuable, if this is indeed appropriate, bearing in mind that not all reviews produce quantitative data for statistical analysis. If you do plan to do this then its wise to invite a suitably experienced and willing colleague to join the review group as it starts to form. This will help their understanding of the review's aims and enable them to suggest relevant statistical techniques that can then be tested during the pilot review.

Our suggestions on forming a review group are summarised in Box 1. Once established, the first task of the group is to develop the review question and protocol and to pilot the review.

1. Determine group size and composition; allowing for different perspectives on the task at hand.
2. Clarify roles and responsibilities of group members, and allow for training opportunities.
3. Try to find resources to support the review.
4. Consider how to ensure an international perspective for the review, by, for example, having an international advisory group.
5. Add new members to the group, with caution.

Box 1 Starting a Review: a summary

Clarifying the review question

A clear review question lies at the heart of all systematic review research. It acts a reference point as the review work progresses and, on completion, indicates to the review audience exactly what has been the quest of the review. It makes sense then that the conclusions reached following synthesis of the primary data included in the review seek to provide answers to the review question. In education systematic review research the review question is unlikely to be of the form that asks for proof that a particular education intervention works or not. Reviews in education seeks

evidence for improvement, answering questions on how to enhance effectiveness, and/or about what works for whom in what circumstances. Wording a review question in this way sometimes produces a long and complicated question and for this reason some review groups decide on aims and objectives for the review. Either is acceptable; one or other is essential. Equally essential is the need for the review team to agree about the question or aims at an early stage. The next stage is to decide how the question or aims will be conceptualized within the wider body of education knowledge.

Conceptual frameworks for BEME systematic reviews

A good report of education research states what conceptual (theoretical) framework guided the research. BEME reviews are no exception and so an important task for the review team is to choose a theoretical orientation to guide the review process. The conceptual framework should be made explicit in the report, as illustrated below. Often the conceptual basis of the review process is implicit. In such cases it is necessary to elucidate from the review's aims and objectives what concepts the reviewers were interested in and how this informed that way the review was conducted, as the following paragraph shows.

Jha et al. (2006) reviewed studies that measured attitudes to professionalism in medicine and the interventions that were effective in changing these attitudes. They write that:

'This review aims to assimilate those studies that report measures of attitudes towards professionalism in medicine. It specifically includes only those studies that report measures that tap into cognitions (i.e. processes that cannot be observed directly). The data elicited from the review will be synthesised to address the following research questions:

- What measures have been used to assess attitudes to medical professionalism?

- What is the psychometric rigour of measures employed to assess attitudes of medical professionalism?
- What interventions have been found to be effective in changing attitudes towards professionalism?' (Jha et al. 2006, p. 823)

From this it is possible to conclude that the reviewers were interested in two concepts: changes in professional attitudes in medical students following interventions *and* the nature and rigour of measures to assess these changes. In contrast, the three examples below show how a particular theoretical model was used by reviewers as an analytical tool and explanatory framework, and the utility of this to show the strengths and weaknesses of the selected theory in a given context. This should not be interpreted as a criticism of the model, rather the role of systematic review results in developing theory in light of empirical findings.

Example One

Milner et al. (2006) report that for their systematic review of the literature regarding clinical nurse educators and research utilization, they used a framework that attempts to describe the complexity of the topic under review. The following quote explains how they used the framework and the reviewers go on to identify how their results did and did not map onto this framework and the utility of this exercise.

'The [review] results are analysed using the revised Promoting Action on Research Implementation in Health Services (PARIHS) framework (Rycroft-Malone et al. 2002). …. We felt it would be useful to use the PARIHS framework to strengthen our analysis and provide insight into its usefulness as a conceptual framework to guide further study in the field. Our findings are discussed in light of its three elements' (Milner et al. 2006, p. 640).

Example Two

The Interprofessional Education review group based their analysis of the impact of an educational intervention on the presage-process-product (3-P) model of

learning and teaching (Hammick et al. 2007). Box 2 explains their choice and the ways in which the model was utilised during the review process. The model resonated with two of the Interprofessional Education review objectives which were to:

- classify the outcomes (products) of interprofessional education and note the influence of context (presage) on particular outcomes, and
- identify and discuss the mechanisms (processes) that underpin and inform positive and negative outcomes of interprofessional education. (p. 736)

Elsewhere (Freeth et al., 2005) we suggested the 3-P model (Biggs 1993, building upon Dunkin & Biddle 1974) as a useful tool for describing and analysing IPE, with utility for putting IPE into practice. The 3-P (presage, process, product) model of learning and teaching was originally devised by Biggs (1993). In his paper, Biggs regarded 'presage factors' as the socio-political context for education and the characteristics of the individuals (planners, teachers and learners) who participate in learning/teaching. 'Process factors' were regarded as the approaches to learning and teaching that were employed in an educational experience and 'product factors' were seen as the outcomes of the learning. Reeves & Freeth (2006) recently applied the 3-P model to the evaluation of an interprofessional initiative in mental health. They found that the model was useful in helping to untangle the complex web of factors that promoted and inhibited success in this initiative. In particular, the model proved effective in yielding new insights, making connections clearer and highlighting the key importance of presage in relation to process and product. Thus the 3-P model served as an analytical framework for the 21 studies and the means of presenting the emergent review findings.
(Adapted from Hammick et al. 2007)

Box 2 Conceptual frameworks: an example

Similarly to Milner et al., Hammick and her colleagues found that mapping review results onto their selected framework was incomplete. This incompleteness of mapping secondary research findings onto a conceptual model highlights the strengths and weakness of a particular theoretical stance in a particular context and where there are gaps in the empirical literature. For example, although Biggs' model includes cost as a presage factor, it is clear that this is a neglected topic for those enquiring into factors influencing the effectiveness of interprofessional education.

Review groups should identify the conceptual base of their review at an early meeting, using it to then inform selection of the primary studies and as a guide during data abstraction, analysis and synthesis. In this way theory acts as a thread connecting different aspects of the review process into a coherent whole. It can also be used to show readers how the different aspects revealed in the analysis and synthesis link with each other: a useful mechanism in the complex world of professional education.

The review protocol

Reeves et al. (2002) write that the protocol provides a detailed description of the review and how you intend to undertake it. In doing so, the protocol should incorporate the following sections: background information that outlines why you are undertaking the review, the review question and/or aims of the review; details of inclusion/exclusion criteria, the search strategy, the methods you will be using for extracting data from studies and details on how you will appraise the quality of the studies that are to be included in the review. The protocol sets out all the initial decisions the review group have taken and provides a mission statement for the group. However, do remember that a review protocol is a living document. It is likely to evolve in the light of subsequent group discussions around issues and problems encountered as you begin the review. The protocol is useful in allowing your peers to scrutinize your intended approach. Such transparency is vital. Peer feedback will identify any potential weaknesses or over-sights in the proposed work. This should strengthen the work and contribute to the production of a more rigorous review. (Adapted from: Reeves et al. 2002)

Once the protocol is completed it should be submitted to BEME for peer review and final approval. You can see examples of protocols on the BEME website **www.bemecollaboration.org**. The following section focuses on the next stage of a review: piloting what is set out in your review protocol.

The pilot review

Piloting the review helps to ensure that the outcomes of your review work are robust and acceptable to the community of practice relevant to the topic under review. Careful planning and attention to detail in the early stages reduce the challenges of the complex review process. Broadly, systematic review work consists of the following major tasks:

I. Collecting and analysing data from primary studies, and

II. Synthesis of secondary data into clear and robust findings and dissemination.

Piloting helps to establish a collaborative group of reviewers who work to agreed and quality tested processes throughout the review. A necessary starting point is a discussion about the types of evidence that might emerge from the review and some of the challenges inherent in secondary analysis of educational research and evaluation. Agreement about what can and cannot be achieved by the systematic appraisal and analysis of primary studies in your topic is essential. The High Fidelity Simulation review group describes the result of their initial discussions as follows:

'Our definition of the elements of a high-quality, systematic literature review is based on previous work by Wolf (2000). The eight elements range from stating the objectives of the review to conducting an exhaustive literature review, tabulating characteristics of eligible studies, synthesizing results of eligible studies, and writing a structured report.' (Issenberg et al. 2005 p. 15).

Pilot review processes

One aim of a pilot review is to ensure that each member of the review team interprets definitions and review processes in the same way. You need to provide the opportunity during the pilot review for the review group to:

- clarify the review question;
- determine data sources;
- consider whether the review question can be answered by the available data.

Retrieval of abstracts and full papers during the piloting of the search strategy provides more data against with key review process can be tested and developed, these include:

- the definition of terms in inclusion and exclusion criteria;
- the data coding sheet;
- meta-analytical and narrative inferences;
- judgements and conclusions.

These lists are not exhaustive. The aim is for well thought through decisions using carefully selected samples of primary studies to establish and agree the working framework of your particular review. You can read more about work done in the early stages of the High Fidelity Simulation review in Appendix 1 (on BEME website **www.bemecollaboration.org**) and Box 3 The pilot process: an examplehas an example of a structured pilot review.

In some reviews the piloting process is informal and implicit, as described below by Veloski et al. (2006):

'During May 2003 a sample of representative articles that had passed the screening by the information scientist and the lead member of the review group was used to pre-test the coding form and instructions to reviewers. Although there were no changes to the overall structure of the form, the wording of many the items was edited for clarification, and several new items were added.' (p. 122, underline added).

This ensured that the processes chosen to select the primary studies for the Feedback and Physician Performance review, and to abstract data from these, processes essential for the work that followed, were sound and fit for purpose.

Pilot I
All review group members reviewed five articles (chosen by the lead reviewer) to determine the scope of the review, to refine the review question, and to assess the applicability of the BEME coding sheet (**www.bemecollaboration.org/**). Following this initial step, we identified areas of the BEME Coding Sheet that required adaptation for our review (e.g. target population; stated intervention; expected learning outcomes; impact of the intervention; and study design); highlighted areas for reviewer training; and further refined the review question. Modifications to the BEME coding sheet were required in most categories.

Pilot II
The second step consisted of a pilot review of 30 articles that addressed all aspects of faculty development (i.e., a focus on teaching, research and administration). Two review group members reviewed each paper, which enabled us to "test" our faculty development BEME Coding Sheet, determine a process for working together, and further refine the review question. At this stage, we decided to focus specifically on faculty development designed to enhance teaching rather than other faculty roles. This step also helped us to finalize our coding sheet, identify additional needs for reviewer training to increase inter-rater reliability, and determine the full scope of the literature search.

(Steinert et al. 2006 pp 499-500)

Box 3 The pilot process: an example

Managing citations, abstracts and primary studies

BEME Guide No 3 described the systematic processes of searching electronically for primary studies (Haig & Dozier 2003). In this section, we discuss issues that arise from mainstream electronic searching, including the matter of translating non-English language papers.

It is likely that together your searches will yield large numbers of studies as has been the case for a number of completed reviews (see completed reviews on website www.bemecollaboration.org). Managing the primary studies well is part of being systematic and transparent: two hallmarks of this type of review work. It also helps with the good practice of being able to easily identify the studies you eventually include in your review and excluded studies, and the reasons for this. The completed

reviews on the BEME website also indicate reasons why studies were excluded from some of the reviews listed. Box 4 lists the items that we recommend review groups report on.

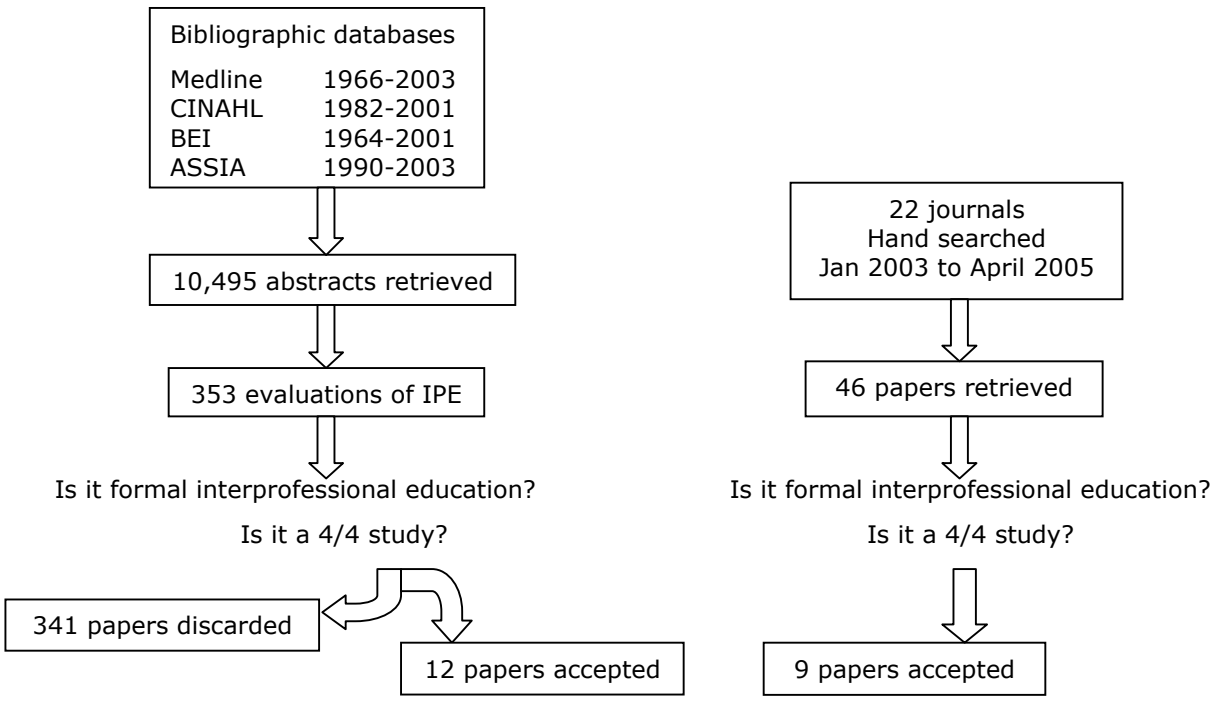
Visual representation of the study selection process is a useful addition to a review report as shown in

Figure 1 for a BEME review. For another example see Evidence based practice in postgraduate healthcare education: A systematic review, by Flores-Mateo G & Argimon JM reported in BioMedCentral, available at **http://www.biomedcentral.com/1472-6963/7/119 in July 2007**.

Numbers of:

- Abstracts found in electronic and hand search process
- Full papers retrieved for matching against review inclusion criteria
- Papers accepted for review & papers rejected (with reasons)
- Papers accepted for review following assessment of study quality & papers rejected (with reasons)

Box 4 Reporting search & selection results: a summary



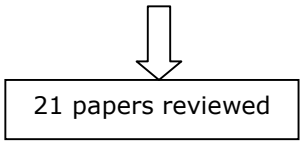


Figure 1 Visual representation of the study selection process for best evidence review of interprofessional education (Hammick et al. 2007 p. 739)

Searching for studies in education is not straightforward and it is sometimes necessary to augment electronic searches with hand searches and citation searching or secondary screening. Box 6 indicates the comprehensive way in which searching for studies was undertaken by the review group who assessed the value of measurements obtained in medical schools in predicting future performance in medical practice. In another review, Issenberg et al. (2005) reported that they:

'… also hand searched key publications that focused on medical education or were known to contain articles on the use of simulation in medical education. These journals included Academic Medicine, Medical Education, Medical Teacher, Teaching and Learning in Medicine, Surgical Endoscopy, and Anesthesia and Analgesia. In addition, we also hand searched the annual Proceedings of the Medicine Meets Virtual Reality Conference and the biannual Ottawa Conference on Medical Education and Assessment. These proceedings include grey literature (e.g., papers presented at professional meetings, doctoral dissertations) determined by our TRG to contain the most relevant references related to our review. Several basic internet searches were also done using the Google.com search engine. The aim was to perform the most thorough literature search possible of peer reviewed publications and reports in the unpublished "grey literature" that have been judged for academic quality'. (p.17)

Extending searches in this way is necessary due to the relative infancy of pedagogic research in health care sciences education and the absence of a tailored taxonomy and an agreed and appropriate descriptive language. This means that it is not possible to totally rely on searching electronic data bases, e.g. Medline, for primary studies that meet your review criteria. Work is in progress to remedy this by establishing new or influencing existing controlled vocabularies, e.g. the Medical Education Taxonomy Research Organisation (METRO). For a full discussion of this, see Haig & Dozier (2003). In the same way, the creation of a taxonomy of topics for indexing medical education by Willett et al. (2008) has the potential to be a valuable

resource for BEME reviews. This ongoing collaborative project can be accessed at http://www.genereauxconsulting.ca/TIMEITEM/application/index.jsp

.. an initial broad scoping search was performed across the key medical and educational databases: ERIC, MEDLINE, Psych Info, Web of Science and Timelit. Significant relevant papers were identified prior to this search, and strategies were drawn up to ensure each of these papers would be retrieved by the scoping search. A series of filtering strategies were developed to remove false hits.

The full search included electronic and non-electronic sources. Multiple Medline searches were conducted and manageable results lists were reviewed. These searches utilised the most appropriate subject headings available, and employed limits to handle very large results sets. The Medline searches were enhanced by searches across other databases, including Embase, Cochrane's EPOC (Effective Practice and Organisation of Care Group) and Controlled Trial databases, and the British Education Index.

The non-electronic search was critical in identifying papers that the databases were unable to realistically produce in manageable quantities. In addition to recommendations from experts, we also conducted hand-searches across key medical education journals: *Medical Teacher*, *Academic Medicine*, *Medical Education* and *Teaching and Learning in Medicine*.

An updating search was conducted to retrieve new research published since the start of the group's work. This search was limited from 2001 to the latest citations and was run across Medline, Embase, Evidence Based Medicine Reviews (including Cochrane), SPORTdiscus, AMED, HMIC, ERIC and BEI. The search strategies used were sensitive, but were not designed for maximum sensitivity, given the impracticality of the massive number of irrelevant citations that would have resulted.

To reinforce the results of all searches, a separate cited reference search was conducted on the Web of Science. Each of the papers included for review from the first search results (as well as several from the updating search) was searched for papers that *cited it* and papers that *it cited*.

(Hamdy et al. 2006 p. 105)

Box 5 Review search processes: an example

Hand-searching

The decision to do a hand search is dependent upon the review topic and is usually made after the electronic searches. Validating the search strategy against a small sample set of articles is considered good practice. Another check can be done through members of the review and advisory group who are likely to be familiar with the literature on the review topic. They may know whether the searches are yielding all the papers possible and of any key grey literature[1] that should be included.

It is useful to estimate how much time hand searching is likely to take before embarking on this task. In some cases it may be a matter of searching journals already searched electronically: Hamdy et al. (2006) comment on this in Box 5. Other reasons include the appearance of a new journal not yet in the usual databases but which is publishing key papers on the review topic as found by Hammick et al. (2007) who commented in their review that:

'Our wider knowledge of the IPE literature, accrued through daily work, alerted us to an additional probable source of good quality studies: Learning in Health and Social Care, a new journal in 2002. This was hand searched from its inception …' (p. 738).

The Early Clinical Experience review group hand searched 'Medical Education, Medical Teacher, Academic Medicine, Teaching and Learning in Medicine, Advances in Health Sciences Education, and the Journal of Educational Psychology. This yielded 21 articles that had not been identified by the main search.' (Dornan et al. 2006 p.5)

Reviews published elsewhere also report hand searching. For example, for the systematic review of studies assessing and facilitating attitudes towards professionalism in medicine, Jha et al. (2007) hand searched three key medical education journals: Academic Medicine, Medical Education and Medical Teacher.' Flores-Mateo G & Argimon JM (2007) report that eight of the 481 'distinctive references' they identified came from hand searches. Milner et al. (2006) state that for their review of research utilization and clinical nurse educators, they

'hand searched the Journal for Nurses in Staff Development, Journal of Continuing Education in Nursing, and Nursing Education Today based on the assumption that relevant articles might have been missed due to different indexing

[1] Materials that cannot be found easily through conventional means, e.g., technical reports, working papers, etc.

approaches within the nursing education community, resulting in additional citations. No additional articles, however, were located.' (p. 641).

Secondary screening

This is another way of ensuring that the search for papers meeting the review criteria is as comprehensive as possible. Examples of this from completed reviews include:

- 'The bibliographies of all articles that fulfilled the inclusion criteria were screened to identify other articles within the review period that fulfilled the inclusion criteria. None were found.' (Dornan et al. 2006 p.6).
- 'The reference lists of existing reviews about medical professionalism were searched for relevant articles.' (Jha et al. 2007 p.824).
- 'We reviewed the reference lists of the relevant original papers and reviews.' (Flores-Mateo G & Argimon JM 2007 p.6).
- 'The results of the database search were augmented by further methods. A cited reference search was conducted on the core papers of relevance examining which papers these cited, and in turn which future papers referred back to the core. Grey literature … searches were also conducted along BEME methodology. Finally, hand searches were conducted across the most relevant journals: Academic Medicine, Medical Teacher, Medical Education, Nurse Education in Practice and Education for Primary Care ... Titles suggesting a focus on self-assessment that had not already been identified were obtained for examination of abstract and if indicated full text. References in full text articles were explored for additional citations.' (Colthart et al. p. 18 2008).

Translations

All reviews are conducted and initially published in the English language. BEME originated from a community of educational practitioners who communicate in English (as the international language of scientific communication) and who work mostly with the English language literature of health care sciences education. BEME aims to publish in other languages and as indicated above there is now one abstract in Spanish and work is in progress to extend this.

Review groups rarely have the resources to pay for the translation of non English language papers. One group that did this reported a cost of 700 Euro for

three papers that were then excluded. However, retrieving non English language abstracts and papers means that review results reflect a truly international perspective and benefit from the work of the international community of health sciences educators. The decision to translate papers needs to balance these positive benefits against the time and finances needed for translation. As we mention above, an international review group has the potential for this to be done by one or more of its members, as was the case with the Faculty Development review. The very few papers in French, Spanish and German found to be eligible were all read by the review group members and translation was not needed. However, no non-English articles were included in the final review. Another group reported that:

'Some non-English language articles were retrieved in the search and were screened when an abstract was available in English. None was selected for inclusion in the review' (Veloski et al. 2006 p.121).

Finally, in this section on the literature a comment on searching for the grey literature. A review group should consider at the outset how important examination of this will be for their topic so that time and resources can be directed appropriately. The grey literature can be a great consumer of time and resources for relatively little benefit: this needs to be considered in light of the planned time for completion and the resources available for all the review processes.

Literature and data management

Systematic reviews yield a lot of data, usually from a large number of primary studies. Sound record keeping is good practice and a useful tool at the report writing stage. Increasingly, BEME review groups use electronic and/or web based forms to maintain the review record and to collect data from the selected primary studies; often linked to commercial products.

Keeping a record of the primary sources of literature

Reporting what literature was and was not included in a review accurately is academic good practice and common sense. You will be dealing with large amounts of literature; some studies are reported in more than one paper and reference lists and bibliographies need to be consistent and easy to proof read. It is essential to have

good administrative support for paper retrieval, tracking, etc. and to ensure that arrangements for this are in place at an early stage.

Using commercial products such as Endnote or Reference Manager and SPSS means you can keep records of primary studies in one place, as the examples in Box 6 show. They enable both electronic and manual checks to be done; and clearly save time.

We developed a local database using Reference Manager™ to maintain citations and abstracts and to review tracking data. This enabled us to use OVID's Direct Export Feature to download the complete record of each citation retrieved by our search into this local Reference Manager database. Ovid's Direct Export Feature is optimised for MEDLINE, but it does not always parse the details of citation records from other databases into the correct fields when downloading to Reference Manager. Therefore, when duplicate citations were available we used either the one from MEDLINE or from the database with the best bibliographic control.

Our initial plan had been to keep both relevant and irrelevant citations in the Reference Manager database. We knew that the software had the capability to identify and purge duplicates when they appeared in later searches. We hoped this would eliminate the need to screen the same citations visually more than once. However, the software was sluggish as the size of the database grew to thousands of records. We retained only relevant citations in the database.
Duplicate citations presented problems throughout the search. As the results of new searches were downloaded, their citations were compared to the Reference Manager database to identify duplicates. Following the recent National Library of Medicine addition of HealthSTAR and Bioethicsline citations to MEDLINE, we encountered many duplicate records in MEDLINE, which Reference Manager was usually able to identify and remove. (Veloski et al. 2006 p.121)

All search results were entered into Endnote (Endnote Version 5.0.2 Research Soft Berkeley, California, USA). Citations from the Ovid databases were saved in a 'Reprint/Medlars' format then imported through the 'Medline Ovid' import filter. Duplicates were discarded, first automatically, then by manual elimination. A first researcher reviewed each of the 6832 articles by title/abstract. She excluded publications that were clearly irrelevant, but retained them in the bibliographic file for future reference.

(Dornan et al. 2006 p.6).

Box 6 Management of References & Citations: examples

Managing data from the primary studies

Accruing data for secondary synthesis involves two processes: abstraction of data and appraisal of the quality of the work being reported. These are often done concurrently as they both demand a close read of the selected primary studies. During this read, as far as is possible the same, relevant data are abstracted from each study, the paper is given a quality score and these data items are then collected together in a systematic way. Forms to do this are either called data abstraction forms (or even sheets), data collection forms or data coding forms. Any one of these term is acceptable, consistency in using the term of choice when reporting a review is essential. In this Guide they are referred to as data coding forms.

We return to quality scores and appraising the worth of the primary studies below; here we look further at data management processes. These must be managed in ways that permit efficient participation in this process by more than one person and allows the data to be effectively drawn upon during the analysis and synthesis of the data. In addition, the data coding form needs to be tailored to the topic of a particular review. The review team needs to set aside time for designing the form to ensure that it is fit for purpose and testing it before using it in the main part of the review. Careful attention to the expected mode of analysis of the data that are collected can save time at the analytical stage.

The BEME website has examples for new groups to follow but it is likely that one of the initial major tasks for review group members will be the development of their review's data coding form. Most reviews develop three categories of data, namely

- Administrative: details of the primary paper including publication and country of origin etc.
- Topic related: including details of the education intervention, types of learners etc.
- Research related: including methodology, data collection methods, analytical approach etc. and the results of the appraisal of the paper.

The data required for administrative purposes are generalisable and this section is likely to look very similar across a number of different reviews. Similarly, data related to the design and methods of investigation used in the primary studies

are likely to be similar whatever the topic under review. This does not apply to the topic related section: to produce a data coding form that is fit for purpose in this respect, it is almost always necessary to pilot preliminary forms. Colthart et al. (2008) report that:

'All members of the review team independently coded a selection of papers into the data abstraction sheets to validate the coding sheets for utility and completeness. All full papers were then read by two group members, using the final version of the coding sheet.' (p. 20).

During the development of the data coding form it is important to focus on the concepts that informed the review question, checking that all aspects of the topic are being accounted for. This is one activity that merits the review group getting together, independently listing data of interest and finally, agreeing on a final data set. This can then be piloted with, say ten primary studies, for completeness (often new items arise from reading the studies) and to ensure there is common understanding of what each data item means.

It is far better to code a little extra data than to have to return to all your primary studies at some time and abstract one (previously neglected) data item from each one of these. This may produce a number of *holes* in the data set, resulting in decisions to disregard incomplete data. This is undoubtedly an easier problem to face than that of searching each study again for one data item.

Once a final data coding form is agreed abstracting data from the complete set of primary studies can either be shared amongst the review group members or assigned to one member with routine quality checks by other members of the group. Sharing distributes the workload and ensures that all reviewers become very familiar with a proportion of the primary studies. Assigning one member or a paid research assistant allows for the development of proficiency in the task and speedier completion. There are not firm rules; each review group chooses what is best for them and puts in place suitable quality control procedures for what is a key phrase of the review process.

Appraising and rating the quality of primary studies

Systematic reviews aim to provide evidence about the effectiveness of an education intervention. This means collecting and analysing data from primary studies that have evaluated the intervention of interest. This is the only way in which the impact of, for example, introducing a new module, curriculum or way of delivering learning can be identified. It allows conclusions about the effectiveness of the intervention in terms of the anticipated outcomes, for example, increased ability to transfer learning to the workplace. It (almost) goes without saying that the quality of secondary research or review is related to the quality of the primary studies. Data used in a systematic review originates from the selected primary studies; their quality determines the quality of the data analysed and synthesised to form the review's conclusions.

Part of the coding process involves making a judgement on the quality of the evaluation using an agreed scale. Box 7 shows how this was done for the review of self assessment and the other tasks done at the same time. The use of the Kirkpatrick model for classifying educational outcomes is discussed later. Note how this review group continued to *search* for additional primary studies at this stage via the reference lists of those already selected.

Box 7 reports how one particular scale quality was used for appraising primary studies; here the focus is on the appropriateness of study design and analysis, and how well the study was executed. Judgements made were then translated into a grade. Note that then studies graded 1 and 2 were excluded from the review. There are many other scales available for this important aspect of secondary research; some of these differ according to the methodology and data collection methods. In particular, there are different criteria for judging studies that have collected quantitative data and those that have collected qualitative data. This may mean having two different versions of the data coding form to take into account the different questions that are asked during the critical appraisal process of the different type of studies. Just occasionally, studies collect both types of data and combining two sections onto one form may help keep data related to one study in the same place. Whatever process of critical appraisal is conducted it is possible to grade or rate each study on a scale similar to that in Box 7. Whoever does this task, the review team need to achieve a consensus about making such judgments and a way of solving any differences. One review group reported that they decided that each

article would be rated by a pair of two reviewers with differences of opinion resolved first in the pair and then if needed, by a third person. They added that, 'in retrospect, the whole process would have been much smoother if only one pair had reviewed all of the articles. A tremendous amount of time was devoted to reconciliation of reviewer discrepancies' (anon).

Obtaining consensus on these processes is an essential part of the quality control processes for any secondary research – more of which later. This is best done during the pilot review as the opportunity for raters to calibrate themselves and come to an agreement on rating items.

Box 7 has a summary of how one BEME review group approached this aspect of review work. Space prevents any further discussion of the many different approaches to critical appraisal of primary research: key principals of this aspect of review work are summarised in Box 8.

Reviewers were asked to rate:

- the appropriateness of the design of the study to answer the research questions posed
- how well the design was implemented
- the appropriateness of the analysis, and to comment on concerns.

They were then asked to comment on what level of the Kirkpatrick Hierarchy (Kirkpatrick, 1967) the outcomes related to. Additionally reviewers identified references cited in these papers that might be of interest to the review and where appropriate these were obtained.

Following data extraction of each paper the two group members independently scored them on a scale of 1 to 5 for the strength of the findings.

Gradings of Strength of Findings of the Paper

- Grade 1 No clear conclusions can be drawn. Not significant.
- Grade 2 Results ambiguous, but there appears to be a trend.
- Grade 3 Conclusions can probably be based on the results.
- Grade 4 Results are clear and very likely to be true.
- Grade 5 Results are unequivocal.

Papers where the conclusions were not supported by the evidence presented i.e. grades 1 and 2 were not considered further.

(Colthart et al. 2008 p.20)

Box 7: A review group's method of judging the quality of primary studies

- Ensure that the review group has discussed why critical appraisal of the primary studies is necessary.
- Choose suitable quality criteria and a rating scale and in a collaborative manner – remember that it is always best to use validated criteria and scales if possible.
- Make sure everyone understands how the criteria and selected scale are operationalised.
- Obtain agreement within the review group on which levels of 'quality' will be included and which omitted from the final set of primary studies.
- Make a note of the reasoning for all decision taken about rating for the final report.
- Identify a practical way of moderating the rating decisions taken by individual members of the review group.
- Read from the list given in Box 9.

Box 8 Choosing and implementing a critical appraisal scale

The BEME website (**www.bemecollaboration.org**) has a list of resources that will help you finalise a robust and transparent way of doing this for your review. Note that both papers by Kruper and her colleagues have useful reference lists for further reading.

Classifying and measuring effectiveness

One aspect of coding that is common to a number of review topics is that of classifying the effectiveness of an intervention according to different educational outcomes. One of the most popular and useful models for this was developed by Kirkpatrick (1967) and has subsequently been tailored to suit specific purposes of other work, including systematic review in education. The Kirkpatrick model has particular value in professional education where the aim is to produce capable practitioners, who can transfer the knowledge, skills and attitudes learnt in the classroom into the workplace and for this to impact upon workplace practices. The original model present education outcomes as:

- learners' reactions;
- learning of skills and knowledge;
- changes in learner behaviour;
- the results of the learning opportunity.

Issenberg et al. (2005) note that

'the Kirkpatrick criteria are nearly identical to Miller's (1990) four-level framework for medical learner assessment (and enables the) effectiveness of medical learning is conceived as an ordinal construct …' (p.15).

They also point out how the effectiveness of learning can be conceived as an ordinal construct ranging from:

- Level 1 - participation in educational experiences
- Level 2a - change of attitudes
- Level 2b - change of knowledge and/or skills
- Level 3 - behavioural change
- Level 4a - changes in professional practice
- Level 4b - benefits to patients

The Kirkpatrick model and its adaptations provide the means of interpreting how individual studies report the outcomes of an intervention and then collecting these into a common format. In the pilot review it is essential to make sure everyone doing this task has the same understanding of this interpretive process. It may be necessary to adapt the model for your particular review topic and add a commentary about each section to show what each means for your review. In Box 9 and Box 10 you can see how two BEME review groups adapted Kirkpatrick's original model for the purpose of classifying outcomes for different educational interventions.

Level 1	**REACTION**	Participants' views on the learning experience, its organization, presentation, content, teaching methods, and quality of instruction.
Level 2A	**LEARNING** - Change in attitudes	Changes in the attitudes or perceptions among participant groups towards teaching and learning.
Level 2B	**LEARNING** - Modification of knowledge or skills	For *knowledge*, this relates to the acquisition of concepts, procedures and principles; for *skills*, this relates to the acquisition of thinking/problem-solving, psychomotor and social skills.
Level 3	**BEHAVIOUR** - Change in behaviours	Documents the transfer of learning to the workplace or willingness of learners to apply new knowledge & skills.
Level 4A	**RESULTS** - Change in the system / organizational practice	Refers to wider changes in the organization, attributable to the educational program.
Level 4B	**RESULTS** - Change among the participants' students, residents or colleagues	Refers to improvement in student or resident learning/performance as a direct result of the educational intervention.

Box 9 The Kirkpatrick model from Steinert et al. (2006 p. 501): an example of adapting for a specific review

Level 1: Participation – covers learners' views on the learning experience, its organisation, presentation, content, teaching methods, and aspects of the instructional organisation, materials, quality of instruction. Level 2: a) Modification of attitudes / perceptions – outcomes relate to changes in the reciprocal attitudes or perceptions between participant groups toward intervention / simulation. b) Modification of knowledge / skills – for knowledge, this relates to the acquisition of concepts, procedures and principles; for skills this relates to the acquisition of thinking / problem-solving, psychomotor and social skills. Level 3: Behavioural change – documents the transfer of learning to the workplace or willingness of learners to apply new knowledge and skills. Level 4:
a) Change in organisational practice – wider changes in the organisational delivery of care, attributable to an educational programme.
b) Benefits to patient / clients – any improvement in the health and well-being of patients / clients as a direct result of an educational programme.

Box 10 The Kirkpatrick model by Tochel et al. (2009): an example of adapting for a specific review

Quality enhancement for BEME systematic reviews

The quality of a BEME systematic review is determined by the experience and capabilities of the review team, conducting the review according to good practice guidelines and using the external expertise to advice and monitor reviews at key stages. BEME quality enhancement processes have much in common with those implemented by organisation doing similar work as Box 11 shows.

Internal review appraisal – Engaging a team of **people with different expertise** to work on and support a review can help ensure both relevance and quality. Those working on a review are often provided with support from advisory groups who provide input at key stages of the review.

External review appraisal – As with many primary research proposals and academic publications, systematic reviews may call on peer-referees to appraise the review at the protocol and final report stage. Indeed, at any point the review it may be beneficial to call on the expertise of relevant external consultants to provide an independent appraisal of the quality and relevance of particular aspects of the review.

Good processes – At various stages in a review, such as **coding** and **appraising** studies, it may be necessary to check these processes are understood and applied consistently amongst the team. At the EPPI-Centre we build in time for moderation exercises with the aim of reducing any ambiguity about how to execute tasks such as coding, before the task begins for real. In addition, it is common practice, and a requirement for EPPI-Centre reviews, that for each study included in a **synthesis** the **data are extracted** by two independent reviewers and their results compared.

From: **http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=1917** accessed 06 Jan 09.

Box 11 Quality enhancement processes for systematic review research in education: an example.

We discussed aspects of forming a review group earlier: the composition of the group is crucial to review quality. As well as their particular expertise, members need enthusiasm for and a commitment to the work to be done. They need to be able to work in a team and willing to learn. Remember that a review group needs a leader and members that have the time to do what they agree to do.

Good practice in BEME review work focuses on undertaking tasks systematically, consistently and reporting this work in ways that are transparent. The development of the review question, search strategy and data coding form are iterative processes that need to be tried and tested. Working in *blind* pairs and having a third person adjudicate any disputed results is good practice. If one person carries out a task setting up a procedure for checking a percentage of their decisions by other review group members helps to ensure that their work is sound. The review group has to combine the need for sound review processes, an awareness of the

state of the art of the literature in their topic and achieving a balance between perfection and practicalities in their approach to quality enhancement. For one review group the research assistant consistently outperformed her seniors, so much so that they could not improve on her article selection by double-screening the citations she identified. They add,

` … much of the literature was of such poor methodological quality that we could not expect perfect inter-rater reliability. We were worried that using strict critical appraisal criteria to arrive at the final dataset would exclude important evidence; it is salutary that some of the strongest evidence was from qualitative research. We used consensus to arrive at our final coding and had to accept that others might not have arrived at the same consensus as us.' (anon).

Finally, BEME recommends that review groups establish an advisory group where possible, or seek external comments from peers on their ideas and work. Peer review also takes place through the BEME process of approving review protocols and scrutiny of final reports.

Training and development for systematic review research

The previous sections have hinted that undertaking a systematic review is a complex and intellectually challenging endeavour. On joining a review group many reviewers find themselves undertaking new tasks such as critical appraisal of education evaluations, making judgements that resonate with the views of others in the review group and making a contribution to reporting the review for a wide audience. As noted above, one of the factors determining the quality of a review is consistency of interpretations and judgements within the review group. Reviewer training can benefit the review work in a number of key areas, including: terminology regarding interventions; the distinction between a research study and a program description with an evaluative component; understanding Kirkpatrick's levels of evaluation; and consensus about research design.

Reviewer training not only develops a greater understanding of the key aspects of secondary research in individual reviewers it also enables a shared and consistent understanding within the group of key concepts and processes. Training can be self-led, for example, review group meetings that aim to achieve an agreed pilot data coding form or they can be led by colleagues with previous BEME review

experience. BEME offers general workshops on systematic review research and workshops tailored to the needs of individual review groups. These usually take place early in the life of the review, resulting in a protocol ready for submission, when data coding is complete with a focus on analysis and synthesis and finally, to guide the report writing stage. Prospective review groups are encouraged to contact members of previous BEME review groups for advice and guidance, especially those groups composed of similar members, i.e. single institution groups, or those without international members.

Coda

This Guide has outlined the process of undertaking a systematic review from the point of the initial idea of a topic to when all data from the selected primary studies has been coded. Its companion Guide No. 14 looks at ways of analysing and synthesising that data, good practice in making judgements from that work and in reporting secondary review research.

Websites of interest

http://www.campbellcollaboration.org/
http://www.cochrane-net.org/openlearning
http://eppi.ioe.ac.uk
http://www.york.ac.uk/inst/crd

References

Biggs J. (1993) From theory to practice: a cognitive systems approach. *Higher Education Research and Development*, 12: 73-85.

Colthart I, Bagnall G, Evans A, Allbutt H, Haig A. Illing J and McKinstry B. (2008) The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice. BEME Guide No 10. *Med Teach* 30(2):124-145.

Dornan T, Littlewood, S Margolis A, Scherpbier A, Spencer J and Ypinazar V. (2006) How can experience in clinical and community settings contribute to early medical education? A BEME systematic review. BEME Guide No 3. *MedTeach* 28(1):3-18.

Dunkin M and Biddle B (1974) *The study of teaching.* New York: Holt Reinhart & Winston.

Freeth D, Hammick M, Reeves S, Koppel I and Barr H. (2005) *Effective Interprofessional Education: Development, Delivery & Evaluation*. Oxford: Blackwell.

Flores-Mateo G and Argimon JM (2007) Evidence based practice in postgraduate healthcare education: A systematic review, *BMC Health Services Research* 2007, 7:119 doi:10.1186/1472-6963-7-119.

Haig A and Dozier M. (2003) Systematic searching for evidence in medical education, BEME Guide No 3. *Med Teach* 25(4):352-363 and 25(5):463-484.

Hamdy H, Prasad M , Anderson M B, Scherpbier A, Williams R, Zwierstra R and Cuddihy H. (2006) BEME systematic review: Predictive values of measurements

obtained in medical schools and future performance in medical practice. *Med Teach* 28(2):103-116.

Hammick M. (2005) Evidence Informed Education in the Health Care Sciences Professions. *Journal of Veterinary Medical Education* 32; 4: 339-403.

Harden RM and Lilley PM (2000) Best evidence medical education: the simple truth. *Med Teach* 22(2):117-119.

Harden RM, Grant J, Buckley G and Hart IR (1999). BEME Guide No 1: Best Evidence Medical Education. Med Teach 21(6): 553-562.

Hart IR and Harden RM (2000) Best evidence medical education (BEME): a plan for action *Med Teach* 22(2):131-135.

Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL and Scalese RJ. (2005) Features and uses of high-fidelity medical simulations that lead to effective learning – a BEME systematic review. *Med Teach* 27(1):10-28.

Jha V, Bekker HL, Duffy SRG and Roberts TE (2007) A systematic review of studies assessing and facilitating attitudes towards professionalism in medicine *Med Educ* 2007; 41: 822–829.

Kirkpatrick D. (1967) *Evaluation of Training. In Craig R, Bittel L, editors. Training and Development Handbook.* New York: McGraw-Hill. p. 131-167.

Miller GE, (1990) The assessment of clinical skills/competence/performance, Acad Med 65 (Suppl. 9), pp. S63-S67.

Milner M, Estabrooks CA and Myrick F (2006) Research utilization and clinical nurse educators: a systematic review, *Journal of Evaluation in Clinical Practice;* 12, 6: 639–655.

Reeves S, Koppel I, Barr H, Freeth D and Hammick M. (2002) Twelve tips for undertaking a systematic review *Med Teach* 24(4):358-63.

Reeves S and Freeth D (2006) Re-examining the evaluation of interprofessional education for community mental health teams with a different lens: understanding presage, process and product factors. *Journal of Psychiatric Mental Health Nursing*; 13: 65-770.

Steinert Y, Mann K, Centeno A, Dolmans D, Spencer J, Gelula M and Prideaux D. (2006) A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No 8. *Med Teach* 28(6):497-526.

Tochel C, Haig A, Hesketh A, Cadzow A, Beggs K, Colthart I and Peacock H. (1999) The Effectiveness of Portfolios for Post-Graduate Assessment and Education: BEME Guide No 12. *Med Teach* 31(4):299-318.

Veloski J, Boex JR, Grasberger MJ, Evans A and Wolfson DB. (2006) Systematic review of the literature on assessment, feedback and physicians' clinical performance *Med Teach* 28(2):117-128.

Willett TG, Marshall KC, Broudo M and Clarke M. (2008) It's about TIME: a general-purpose taxonomy of subjects in medical education. Med Educ 42(4):432-438.

Wolf FM. (2000) Lessons to be learned from evidence-based medicine: practice and promise of evidence-based education, Med Teach 22:251-259.