Evaluation Methodologies

A brief review of Meta-evaluation, Systematic Review and Synthesis Evaluation methodologies and their applicability to complex evaluations within the context of international development

Date //June 2011

By// Kristin Olsen and Sheelagh O'Reilly

IOD PARC is the trading name of International Organisation Development Ltd//

Omega Court
362 Cemetery Road
Sheffield
S11 8FT
United Kingdom

Tel: +44 (0) 114 267 3620
www.iodparc.com

Introduction

This review provides a brief overview of 3 synthesis methodologies used to deliver wider learning (what works where and why) around evaluation processes and specified topics: meta-evaluation, synthesis evaluations and systematic reviews. Each of these processes can be used to review and analyze data from a range of sources and provide a mechanism to draw out learning through the use of an explicit or implicit protocol specifying key details of the process including data quality assessment.

This paper seeks to describe each methodology and their key characteristics, and provide an assessment of their applicability for use in the evaluation of complex interventions or initiatives in the international development context.

1. Meta-evaluation

1.1 Meta-evaluation is a systematic and formal evaluation of evaluations, evaluation systems or use of specific evaluation tools in order to guide planning / management of evaluations within organisations. A meta-evaluation can be used for ongoing evaluations (formative) or report on the strengths and weaknesses of previous evaluations (summative). It was a methodology proposed by Michael Scriven in 1969 to describe his evaluation of a plan to evaluate educational products (Scriven, 2009).

Stufflebeam (2000) describes the meta-evaluation as "the process of delineating, obtaining, and applying descriptive information and judgmental information - about the utility, feasibility, propriety, and accuracy of an evaluation and its systematic nature, competent conduct, integrity/honesty, respectfulness, and social responsibility - to guide the evaluation and/or report its strengths and weaknesses"

1.2 A meta-evaluation is carried out by making an assessment of evaluations through reports and other relevant sources including information, and judgements, from stakeholders including the evaluator, client, programme staff, programme beneficiaries, and others (Stufflebeam, 2001; 2011). The assessment is based upon a set of critically recognised evaluation standards or check list[1] such as the American Evaluation Association's Programme Evaluation Standards (AEA, 1994), adapted for purpose. In practice a wide range of criteria are used in meta-evaluations ranging from predetermined and structured (such as AEA Programme Evaluation Standards) to emergent and unstructured. The approach used in applying criteria often reflecting the meta-evaluation procedure itself (Cooksy and Caracelli, 2009) –from narrative reviews to simple audits.

Meta-evaluation is widely used in international development, including by DFID (Perrin, 2009), ALNAP[2] (ALNAP, 2003) and the UNICEF Evaluation Office. Meta-evaluations considered for the purposes of this review applied standard evaluation criteria adapted to incorporate internal standards and reflect the needs and principles of the audience/ organisations. For example ALNAP uses a Quality Proforma (QP) developed specifically for meta-evaluation of humanitarian action evaluations including a section to assess the humanitarian action-specific 'Analysis of Intervention'. The ALNAP QP aims to be a live and evolving document, and is being applied by a number of other agencies including the UK Disasters Emergency Committee and Groupe URD.

[1] For example the Program Evaluations Metaevaluation Check list developed by Stufflebeam in 1999 available online: www.mattburnell.com/clients/wmu/archive_checklists/program_metaeval.pdf
[2] Active Learning Network in Accountability and Performance in Humanitarian Action

In international development meta-evaluations the assessment frameworks include a qualitative rating of the criteria (such as good, satisfactory, unsatisfactory and poor) as well as a narrative analysis targeted at key stakeholders and with additional emphasis on extracting learning and good practice, and wider learning beyond the organisation. An assessment of how and whether the original evaluations themselves applied standard OECD-DAC evaluation criteria are also important in this context. In addition ALNAP have adapted their methodology to increase participation of stakeholders (including programme staff and evaluators) in the meta-evaluation process. The Global Evaluation Report Oversight System (GEROS) developed by IOD PARC for UNICEF Evaluation Office describes an extensive protocol that enables consistent analysis of reports, including requirements that all evaluators undergo training in using the protocol and provide targeted reporting according to the audience (including senior management) (Barnes et al, 2010).

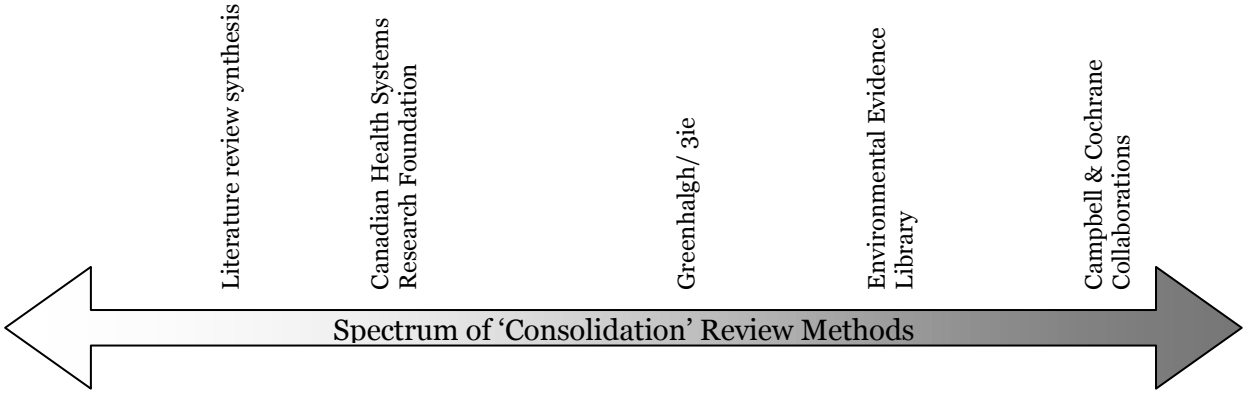
Key characteristics of a meta-evaluation include:

Issue	Key Characteristic
Design	Design of a protocol that includes sampling procedures, an assessment and analysis framework, and specifying reporting requirements. The assessment framework is developed based on the application of a set of critically recognised standards e.g. American Evaluation Association's Programme Evaluation Standards (Utility standards[3] Feasibility Standards[4], Propriety Standards[5] and Accuracy Standards[6]) adapted to suit the meta-evaluation in question. Meta-evaluations in the realm of international development include criteria around the appropriate use of OECD-DAC criteria including effectiveness, efficiency, relevance, sustainability, impact and additional humanitarian criteria (coverage, coordination, coherence and protection).
Sampling	Meta-evaluations can be applied to a single or multiple evaluations and tend to focus around a topic or type of programme, or evaluations carried out by an organisation within a certain timeframe (e.g. 1 year). Samples are taken from evaluation reports, and sometimes from stakeholders. Depending upon the purpose (and scope) of the meta-evaluation sampling can be deliberative, random or include all reports available. Inclusion criteria can be applied.
Data collection	Data is sourced from evaluation reports, and may include supporting information sourced through stakeholders through field visits or interviews.
Analysis	Reports are reviewed using the agreed systematised protocol or guidelines. Examples include the ALNAP QP (ALNAP, 2003) and UNICEF Global Evaluation Report Oversight System (GEROS) (Barnes et al, 2010). Analysis includes ratings and narrative assessment around criteria, identifying trends, strengths and weaknesses to enable enhancement of future evaluation practice and provide guidance on the credibility or *confidence to act* on the findings of the evaluation.

[3] The utility standards are intended to increase the extent to which program stakeholders find evaluation processes and products valuable in meeting their needs.
[4] The feasibility standards are intended to increase evaluation effectiveness and efficiency.
[5] The propriety standards support what is proper, fair, legal, right and just in evaluations.
[6] The accuracy standards are intended to increase the dependability and truthfulness of evaluation representations, propositions, and findings, especially those that support interpretations and judgments about quality.

2. Systematic Review and Synthesis Evaluation: a Spectrum?

Broadly speaking both systematic reviews and synthesis evaluations seek to provide evidence around a topic through synthesising what works, where, how and why. Whereas Cochrane reviews at one end of the 'spectrum' draw largely on experimental reviews (such as Randomised Control Trials - RCTs), further across the spectrum (towards more qualitative synthesis evaluations) systematic reviews include a range of data sources (both quantitative and qualitative) and employ a number of review methods, including meta-ethnography, narrative synthesis, realist synthesis and thematic synthesis for qualitative data as well as the meta-analysis approach to quantitative data (3ie). Along this 'spectrum' of synthesis evaluations are those supported by the Environmental Evidence Library[7] that promotes and disseminates systematic reviews of evidence on the effectiveness of human interventions in environmental management and the environmental impacts of human activities drawing on some experimental evidence. Further along the spectrum Greenhalgh (2004) brings together material from a wide variety of sources and applies an 'iterative and pluralist approach' to defining evidence (meta-narrative mapping) on the basis of criteria such as scholarship, comprehensiveness, and contribution to subsequent works'. Whilst a systematic review is a form of synthesis evaluation – a synthesis evaluation is more likely (but not necessarily) to focus on a broader topic (rather than a specific research question), is less likely to draw on experimental evidence and will have a less formal protocol that may not be peer-reviewed.



The diagram is an informal initial attempt to place different approaches to synthesis or 'consolidation' on a single spectrum. Synthesis reviews and systematic reviews, with all the variants between are attempts to, with different degrees of formality and using different types of evidence, to 'consolidate' i.e. draw together the results of research activities into a 'current state of the art' document that can be used by other experts and policy makers.

2a. Systematic Review

2.1 A systematic review is the process of methodologically identifying, critically appraising, and selecting (high quality) evidence relevant to a specific research question and synthesising the results. Systematic reviews have principally been applied in the health sector, particularly systematic reviews of randomized control trials for evidence-based medicine. Two principal proponents of the systematic review include the Cochrane Collaboration[8] and the Campbell Collaboration[9]. The Cochrane Collaboration is a research network aiming to create a high quality evidence-base for human health care through preparing, updating and promoting access to Cochrane reviews. The Campbell Collaboration is a similar network dedicated to promotion of systematic reviews in the social and behavioural sciences through the Campbell reviews.

[7] www.environmentalevidence.org/Library.htm
[8] www.cochrane.org/cochrane-reviews
9 www.campbellcollaboration.org/systematic_reviews/index.php

2.2 Systematic reviews are characterised by the rigorous application of a peer-reviewed protocol, peer involvement in identifying criteria for selecting and including/ excluding evidence for review, systematically extracting and organising data, and use of rigorous analysis methods. Also here protocols are adapted according to their application. The process for conducting a systematic review is similar however and is summarised below (Higgins and Green, 2009; Campbell Collaboration, 2001; Greenhalgh, 2004; Stewart et al, 2010):

Process for a Systematic Review
a) Formulate review/ research question b) Develop a detailed protocol for review. This is an explicit, rigorous, transparent and replicable procedure for the review c) Identify studies to be included in the review according to carefully selected, reproducible inclusion and exclusion criteria d) Locate relevant research/ studies through a search strategy e) Select studies to be included in the review according to the inclusion and exclusion criteria f) Analyse the quality of the study according to a quality appraisal strategy that is relevant to the review question and the types of studies under review (including appraisal of primary research designs and study characteristics) g) Extract data using systemised data extraction forms that enable synthesis and analysis h) Synthesize studies and apply systematic coding using methods appropriate to the review type and type of data available i) Analyse included studies using appropriate methods j) Interpret results

Systematic reviews are also increasingly common in international development, particularly in the context of evaluating impact. Advocates within the sector also lie along a 'spectrum' according to the types of data that are used. For example the Abdul Latif Jameel Poverty Action Lab (J-PAL) focus on Randomized Evaluations (REs)[10] to answer questions critical to poverty alleviation towards delivering a scientific evidence base for policy makers. Further across this spectrum is the work of the International Initiative for Impact Evaluation (3ie) which has partnered with the Campbell Collaboration to produce systematic reviews in accordance with Campbell Collaboration guidelines and procedures. 3ie are currently funding (together with DFID and AusAID) a large number of new systematic reviews in international development to address key issues for example: What is the impact of initiatives to reduce sexual and gender-based violence in conflict and post-conflict situations?(3ie).

Other recent examples of systematic reviews funded by DFID include a systematic review to discover the impact of microfinance on poor people in order to address concerns that this is not the case (Stewart et al, 2010) and a systematic review of interventions to promote social cohesion in sub-Saharan Africa (Kinga et al, 2010). Through a rigorous and extensive process of searching and screening for relevant studies, each of these reviews identified a number of studies (15 for Stewart et al; 8 for Kinga et al) that addressed the research questions and were considered rigorous enough. Studies considered by Stewart et al included four randomised controlled trials, two non-randomised controlled trials and nine case control studies. Kinga et al included studies with randomized treatment assignment or clear, quasi-experimental delineation of treatment and comparison groups, as well as pre- and post-intervention measurement.

[10] A Randomized Evaluation is a type of impact evaluation that uses random assignment to allocate resources, run programs, or apply policies as part of the study design.

Each of the studies succeeded in drawing important conclusions with regards the impact and understanding of the interventions on development outcomes. However each study also encountered significant challenges (limiting certain areas of analysis) with the availability and quality of data and with differences in intervention design and measurement. Despite being aware of the complexity of microfinance as a development intervention, Stewart et al (2010) encountered greater variety than anticipated and a resulting lack of consistency in both the interventions evaluated and the outcomes measured. In addition the review struggled to define and measure important qualitative outcomes such as empowerment, which lacked 'standard succinct' outcome measures. Further the studies provided varying levels of detail around the interventions themselves including information around possible contamination due to other microfinance programmes. Similarly Kinga et al (2010) found that the limited number of studies and lack of information and analysis of contextual factors limited the extent to which context could be considered within the causal analysis. The y also highlighted the ongoing challenge of determining measures of social cohesion effects that are reduced sufficiently to provide useful measurement yet specific enough to enable comparison across contexts.

Each review included strong recommendations that further study (including research into validity and comparability across contexts and more complementary qualitative studies) was required. This suggests that results were not entirely conclusive and would benefit from being considered in light of further work.

2b. Synthesis Evaluation

2.3 A synthesis evaluation is a formal approach (usually expert-led) of drawing together literature and studies on a specific topic with the purpose of providing analysis.

2.4 Synthesis evaluations are also widely applied in different sectors including in international development. They include literature reviews such as Liverani and Lundgren's (2007) analysis of evaluation systems in Development Agencies 11 as well as more complicated reviews such as the expert-recommended Decision Support Synthesis (DSS) approach developed by the Canadian Health Systems Research Foundation (Lomas, 2007), through a collaborative project with the NHS in the UK12. This latter approach specifically seeks to synthesise evidence (including but not limited to research) for decision-making thereby assuming that the synthesis requires judgment and interpretation, rather than just the ability to summarise. This process is a collaboration between decision-makers engaged in defining the scope and research question, and making recommendations for management and policy, and researchers who are engaged in each stage of the process.

2.5 The process is summarised below (Lomas, 2007):

Process for a Synthesis Evaluation
a) Defining the scope of the review b) Summarizing research c) Drawing out implications from research d) Creating recommendations for management and policy

[11] Literature review of OECD-DAC peer reviews formed the basis of Liverani and Lundgren's (2007) analysis of evaluation systems in Development Agencies

12 National Health Service (NHS) Research and Development program on the Methods of Synthesis for Managers and Policy Makers project

The ongoing complex evaluation of the Paris Declaration (PDE) commissioned by the Paris Declaration Evaluation Secretariat is a mixture of synthesis methods (literature review and a synthesis of different country evaluations). It is a combination of a meta-evaluation that synthesises the evidence from a number of evaluations at the country level which have used a similar protocol, and combines this evidence (through a synthesis process along the 'centre of the spectrum' – see diagram above) with the studies carried out by donors and other agencies plus literature review and other evidence (IOD PARC, 2009). It demonstrates how different tools can be combined to provide important evidence in a highly complex and contested environment. The PDE itself was in a position to commission the evaluations based around an agreed protocol and therefore the format and style of evidence that would feed into the overall synthesis. Although there is a tendency to use this approach for joint evaluation this approach is still somewhat unusual as most meta-evaluations and most synthesis work use pre-existing material. The work inside the PDE is a true mixed methods approach combining formality of a protocol for the country evaluations to facilitate some comparisons with the flexibility of a synthesis process that enables quality of evidence to be considered as well. [13]

2.6 Key characteristics of systematic reviews and synthesis evaluations include:

Issue	Systematic Review	Synthesis Evaluation
Design	A peer-reviewed and pre-determined process from identifying research question and the research protocol, analysis and interpretation of results	A formal design is usually developed e.g. using the Canadian Health Services Research Foundation approach[14], tailored to the needs of the review. The design is ideally expert-led. However synthesis reviews at their lightest can be relatively informal literature reviews of accessible material.
Sampling	Peer review group help to identify studies (studies identified e.g. through electronic databases, hand-searching key journals, pursuing references of references, and advice from experts) according to standards. For example the Cochrane Review accepts mainly original studies of randomly controlled trials and clinically controlled trials; the Campbell Review (and therefore method proposed by 3ie includes qualitative research, grey literature and unpublished studies). Studies to be considered are selected on the basis of inclusion criteria e.g. technical and sectoral relevance with regards the research question as well as criteria relevant to design and quality of primary research to enable analysis and comparison.	Samples for review generally include existing studies and reports sourced through the experts and clients involved. Explicit standards are applied to ensure that evidence used is grounded in evidence and research. Decisions around inclusion of studies is guided by criteria established on the basis of the purpose of the evaluation e.g. ability to respond to evaluation questions

[13] The PDE evaluation is not yet public. When it is it will contain a Technical Annex that will address these issues in detail.
[14] Often used by ODI for synthesis reviews.

Issue	Systematic Review	Synthesis Evaluation
Data collection	Data mainly extracted from experimental data (such as RCTs), using data extraction forms. Qualitative data is coded according to key characteristics such as research question, research design, validity, robustness of methods, sample size, power, nature and strength of findings, validity of conclusions (Greenhalgh, 2004), outcome measures. Cochrane Review places data sourced through RCT at the top of a 'hierarchy of evidence' quality in research design and potential for high internal validity.	Data is extracted from a range of different sources including programme reports, qualitative and quantitative research, grey literature and unpublished studies. Data can come from experimental focused studies as well as from qualitative sources.
Analysis	Use of rigorous analysis methods. Cochrane Review requires statistically rigorous meta-analysis. Greenhalgh's narrative (rather than statistics) can be used to synthesise data to overcome issue of different researchers' subjectivity, methods, approaches etc – and generating systematic information that could be analysed.	Analysis is more likely to be narrative and systematically analysed around the review questions. Analysis should include assessments of data quality.

3. Applicability of methodologies for use in complex evaluations in international development, and to informing robust evidence-based policy making

Relevance of methodologies

Each of these methodologies provide mechanisms for utilising evidence from multiple sources to provide coherent conclusions on that evidence base and therefore give some degree of measurable confidence to the recommendations that are made. However the purpose of a meta-evaluation, and therefore the role it plays within complex evaluation processes differs significantly from a systematic review and a synthesis evaluation.

A meta-evaluation is used for understanding the evaluation process within an organisation or around a specific topic and as such generates an understanding of how the evaluation process itself can contribute to a better understanding of aid effectiveness (through recommendations produced by reliable evaluations). Meta-evaluations can help to identify systematic weaknesses in the way that an organisation approaches evaluation which may be compromising their ability, with their partners, to improve aid effectiveness. Meta-evaluation standards are adaptable in a transparent manner to changes in institutional, national and international requirements around development practice.

Systematic reviews and synthesis evaluations/analyses on the other hand try to synthesise what works, where, how and why, with consideration to the quality of the evidence. A systematic review can provide reliable, robust and unbiased evidence with regards a specific research/ policy question. However the relevance of the evidence is highly specific to the context which the review was able to consider, which is in turn limited by the availability of quality experimental data within the international development context. Synthesis evaluations draw on a broader range of data and information sources that are available and are less likely to have a formal peer reviewed protocol.

Evidence suggests that the more complex the intervention – the greater the need for a range of different evaluation tools and processes to identify and explain the impact of interventions – as well as to understand how to communicate findings and go forward in terms of promoting positive impacts (Rogers and Watts 2008; Rogers 2008). It may be that a combination of different synthesis methods may be appropriate. One consideration is the extent to which 'comparative analysis' is to be undertaken and how far the approach can control e.g. at country level for different contexts. Formal methods do exist for qualitative comparative analysis but at present these have not been used in the field of international development and would require significant skills to be able to utilise this approach including strong initial design.[15]

Complexity of Development Interventions
Within the arena of international development many initiatives do not have simple causal relations between inputs/outputs and hence outcomes. These interventions into existing social-economic-political relations will achieve outcomes and impact through a range of mechanisms not all of which are transparent in the initial analysis. Within any single initiative the way that results and impact are achieved may well vary at different stages throughout the intervention – possibly over a long period of time, when external conditions are also changing. For instance the processes to assess the ability of different rice varieties to grow in context requires direct evaluation methods e.g. RCTs, but it is likely that the process of distribution, including setting up new approaches e.g. community based systems will be complicated although the key logic of the situation will be relatively clear. However assessing the impact in complex social-cultural situations with differential land holdings is a complex process and likely to call for significant contextual

[15] See recent review of Charles Ragins new work 'Redesinging Social Inquiry: Fuzzy Sets and Beyond. Vaisey, Stephen (2009). QCA3.0: The "Ragin Revolution" Continues. *Contemporary Sociology: A Journal of Reviews* **38**: 308

knowledge during the evaluation. Similar initiatives may take place in different social-cultural situations and evidence of impact may both vary and be assessed using different processes and tools. Understanding the way that that the complexity of the situation affects what works, where, why and how can be facilitated by the use of an appropriate synthesis tool.

As well as considering the relevance of the tool according to the purpose of the synthesis the applicability of the 3 methodologies in delivering evidence for influencing policy can be considered in light of key factors that influence use and uptake of evaluation evidence by policy and decision makers. This includes quality and accuracy, credibility, timeliness and practicality and accessibility (Hyjek, 2010) as well as considerations of cost-effectiveness.

Quality and accuracy of evidence

In their very nature meta-evaluations have a key role to play in assessing the quality and accuracy of evaluations themselves. The quality and accuracy of the meta-evaluation requires that the assessment is carried out on the basis of critically (internationally) approved criteria that are agreed if possible with appropriate stakeholders and applied using clear replicable guidelines. As one is reviewing the quality of evaluations and the processes under which they were commissioned and undertaken this approach will not, directly link to the topics of the evaluation under review unless this was in the initial protocol design.

Systematic reviews aim to produce high quality evidence through reducing bias by systematically searching all of the literature and extracting relevant quality evidence (3ie). Provided there is an effective [unbiased] peer review process bias can be reduced. However there are still issues around the availability of quality experimental evidence (Stewart et al, 2010; Kinga et al, 2010) and the 'scale' at which the evidence was obtained. For example relatively small pilot activities might well obscure issues that could arise when the intervention is 'scaled up' to a wider group or into normal government systems (White, 2000).

The strength of synthesis evaluations are in their ability to triangulate through drawing on a range of data sources. However the analyses should include an assessment of the data sources used, and therefore of the evidence provided.

Credibility

The credibility of evidence is linked to the credibility of its source e.g. a respected expert or organisation, and in terms of its fit with professional and practice wisdom (Hyjek, 2010; Nutley et al, 2007). Evidence generated from systematic reviews is therefore likely to be considered credible, as they often draw on the expertise of known academics and/ or institutions and are funded through donor agencies. In addition to the above, credibility is also likely to be increased through transparent processes such as the process of developing and publishing the protocol for systematic reviews and carefully documenting its progress making it possible to scrutinize the methods and audit the process. A systematic review is also more rigorous than a literature review as anyone could follow the review protocol and (hopefully) arrive at similar conclusions (3ie) i.e. this follows the idea of publically available experimental design which allows replication by others.

Timeliness and Practicality

Policy and decision makers require information to be available within an appropriate timeframe, as well as providing conclusive answers. Clearly information is more readily available where the data already exists i.e. where syntheses draw on existing studies. In this context synthesis reviews based on new research provide more long-term benefits as discussed above. Arguably more complex reviews (such as synthesis reviews) and use of multiple synthesis tools take more time, but provide more robust evidence rather than a reliance on one or two 'stand alone' studies. Synthesis reviews do not

lead to quick answers, but the robust nature of the output can increase the credibility of the evidence, especially if it contradicts 'received wisdom'.

Accessibility

The presentation of findings and evidence is important to policy and decision makers who can be put off by academic style reporting or complex analyses (Hyjek, 2010; Nutley et al, 2007). Each of the methodologies considered have tendencies to be academic exercises removed from stakeholders, although each has the scope for increasing participation of target audiences during the process. The meta-evaluation should where possible be an interactive exercise where dialogue on evaluation quality and process between the meta-evaluators and stakeholders can take place – creating dialogue around lesson learning and good practice, and enabling better use of meta-evaluation findings. Synthesis evaluations have the scope to involve peers in the process of defining the research question as well as in interpreting results. Systematic reviews are often highly academic exercises and are less likely to foster a sense of 'ownership'. However ensuring that peer groups are representative of a range of disciplines and target audiences could help to address this. Communication of evidence to non-experts is key to increasing accessibility of evidence, and its strengths and weaknesses – both during the review process and at its conclusion. Ownership is likely to be a factor influencing accessibility of the evidence.

Cost-effectiveness

Meta-evaluations can be relatively cost-effective as they can be conducted through desk reviews and telephone interviews. Costs are higher when they include field visits to consult directly with stakeholders. There is rarely a budget provided for either carrying out meta-evaluations or follow-up e.g. targeted feedback however.

Systematic reviews can be expensive processes as a large amount of resources (including technical capacity) are required to deliver answers to a single research question. Review of existing information is relatively cost-effective however implementing new studies e.g. RCTs can be expensive and must ensure key factors such as political will and other enabling conditions and that information can not be gained from other sources. However in the long-term good systematic reviews can help to identify gaps in the knowledge to help shape future research agendas and reduce duplication of research by making it clear what we do know and what we need to know, as well as reducing the need for traditional narrative reviews that tend to be produced over and over again (3ie).

References

ALNAP (2003) ALNAP Review of Humanitarian Action in 2003: Improving monitoring to enhance accountability and learning, Chapter 4 Meta-evaluation, available online: http://www.alnap.org/pool/files/FCh401bp.pdf

AEA (1994) American Evaluation Association Programme Evaluation Standards, www.eval.org/evaluationdocuments/progeval.html

Balthsar, A (2006) The Effects of Institutional Design on the Utilization of Evaluation: Evidenced Using Qualitative Comparative Analysis (QCA), *Evaluation*, Vol 12(3): 353–371

Barnes, J, Noij, F, Watson, S, Struwig, A, Gayfer, J and Dinsmore, H (2010) UNICEF Global Evaluation Report Oversight System Assessment Methodology, IOD PARC methodology developed for UNICEF Evaluation Office, *Unpublished*

Brass, CT, Nunez-Neto, B. and Williams, ED. (2006) Congress and Program Evaluation: An Overview of Randomized Controlled Trials (RCTs) and Related Issues. CRS Report for Congress. http://digital.library.unt.edu/ark:/67531/metacrs9145/

The Campbell Collaboration (2001), *Campbell Systematic Reviews: Guidelines for the Preparation of Review Protocols*, available online: www.campbellcollaboration.org/artman2/uploads/1/C2_Protocols_guidelines_v1.pdf

Chalmers, I (2005) If evidence‐informed policy works in practice, does it matter if it doesn't work in theory? *Evidence & Policy: A Journal of Research, Debate and Practice*, Vol.1 (2): 227‐242

Cooksy, LJ and Caracelli, VJ (2009) Metaevaluation in Practice: Selection and Application of Criteria, *Journal of MultiDisciplinary Evaluation*, Vol.6 (11) http://digital.library.unt.edu/ark:/67531/metacrs9145/

European Evaluation Society (2007) EES Statement: The importance of a methodologically diverse approach to impact evaluation – specifically with respect to development aid and development interventions . http://www.europeanevaluation.org/images/file/Member_section/EES%20Statement%20on%20Methodological%20Diversity/EES%20Statement.pdf

Glouberman, S. and B. Zimmerman (2002) *Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like? Commission on the Future of Health Care in Canada,* Discussion Paper 8. Available at www.hc-sc.gc.ca/english/pdf/romanow/pdfs/8_Glouberman_E.pdf

Greenhalgh, T, Robert, G and Bate, P. (2004) How to Spread Good Ideas: A systematic review of the literature on diffusion, dissemination and sustainability of innovations in health service delivery and organisation. Report for the National Co-ordinating Centre for NHS Delivery and Organisation R & D (NCCSDO). http://www.sdo.nihr.ac.uk/files/project/38-final-report.pdf

Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2009. Available from www.cochrane-handbook.org.

Hyjek, J (2010) Lesson Learning/ Promoting Uptake and Use of Evaluation Evidence by Policy and Decision Makers, IOD PARC Report for Evaluation Department, DFID, available online: http://iodparc.com/resource/promoting_uptake.html

IOD PARC (2009) Technical Proposal for the Joint Evaluation of the Paris Declaration, Phase II, Core Team, PDE17, *Unpublished*

International Initiative for Impact Evaluation (3ie), www.3ieimpact.org/systematicreviews/systematic_review.php

Kinga, E, Samiib, C and Snilstveitc, B (2010) Interventions to promote social cohesion in sub-Saharan Africa, *Journal of Development Effectiveness*, Vol.2 (3): 336–370

Liverani, A and Lundgren, HE (2007) Evaluation Systems in Development Aid Agencies: An analysis of DAC Peer Reviews 1996-2004, *Evaluation* 13(2): 241-256

Lomas, J. (2007) CHSRF Knowledge Transfer: Decision Support: A New Approach to Making the Best Healthcare Management and Policy Choices, *Healthcare Quarterly*, 10(3): 16-18

Nutley, S, Walter, I and Davies, H (2007) Using Evidence. How Can Research Inform Public Services, Bristol: The Policy Press

Perrin, B (2009) Review of the Quality of DFID Evaluations: A Delicate Balancing Act, Report to the Independent Advisory Committee on Development Impact (IACDI)

Rogers, PJ (2008) Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions, *Evaluation*, Vol. 14 (1): 29 – 48 Available http://evi.sagepub.com/cgi/content/abstract/14/1/29

Rogers, P and Watts, J (2008) Methodological Innovations in Impact Assessment of Agricultural Research, International Workshop on Institutional Learning and Change Initiative, Brasilia, Brazil http://www.cgiar-ilac.org/content/presentations-ilac-team

Scriven, M. (2007). *Key evaluation checklist*. Retrieved January 22, 2009, from the Western Michigan University Web site: http://www.wmich.edu/evalctr/checklists/kec_feb07.pdf

Scriven, M (2009) Meta-evaluation Revisited, *Journal of Multidisciplinary Evaluation*, Vol.6 (11) http://survey.ate.wmich.edu/jmde/index.php/jmde_1/article/view/220/215

Stewart R, van Rooyen C, Dickson K, Majoro M, de Wet T. (2010) What is the impact of microfinance on poor people? A systematic review of evidence from sub-Saharan Africa (Technical report). London: EPPI-Centre, Social Science Research Unit, University of London http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2965&language=en-US

Stufflebeam, DL (2000) The Methodology of Metaevaluation as Reflected in Metaevaluations by the Western Michigan University Evaluation Center, *Journal of Personnel Evaluation in Education*, Vol.14 (1): 95-125

Stufflebeam, D (2001) 'Guiding Principles Checklist for Evaluating Evaluations' http://www.wmich.edu/evalctr/checklists/guiding_principles.pdf

Stufflebeam, D. (2011) Meta-Evaluation. *Journal of MultiDisciplinary Evaluation* [Online] Vol.7 (15) Available: http://survey.ate.wmich.edu/jmde/index.php/jmde_1/article/view/300/295

Vaisey, Stephen (2009). QCA3.0: The "Ragin Revolution" Continues. Contemporary Sociology: A Journal of Reviews 38: 308

White H (2011) An introduction to the use of randomized control trials to evaluate development interventions. *International Initiative for Impact Evaluation*, Working Paper 9 http://www.3ieimpact.org/admin/pdfs_papers/White_RCT_WP%209.pdf