Meta-evaluation of the Impact and Legacy of the London 2012 Olympic Games and Paralympic Games

Developing Methods Paper

INTERIM REPORT – November 2012

For:
Economic and Social Research Council &
Department for Culture, Media and Sport

Prepared by:
Professor David Gough, Institute of Education, University of London
Professor Steve Martin, Cardiff Business School
Ecorys
Grant Thornton UK LLP

November 2012

Contents

1 Introduction

This report brings together the findings from phase one of the Developing Meta-Evaluation Methods study, which is being undertaken in conjunction with the Meta-Evaluation of the Impacts and Legacy of the London 2012 Olympic Games and Paralympic Games. The Meta-Evaluation has been commissioned by the Department of Culture, Media and Sport (DCMS). The work on methods is funded by the Economic and Social Research Council (ESRC)[1]. The aim of this element of the study is to review and advance understanding of methods of meta-evaluation.

1.1 Background

In May 2010, Grant Thornton, ECOTEC Research and Consulting (now Ecorys) and associates were commissioned by the UK Department for Culture, Media and Sport (DCMS) to conduct a comprehensive three-year Meta-Evaluation of the Impacts and Legacy of the 2012 Olympic Games and Paralympic Games. The study is of the utmost importance in demonstrating the legacy impact of the 2012 Games across all thematic areas and will be the single largest and most comprehensive evaluation exercise commissioned in connection with the event. The study will involve:

"… the synthesis of results, findings and the outputs across a set of existing and planned evaluations with heterogeneous features, into a single overall evaluation. It will also involve reviewing the methodology of the project level evaluations to assess whether they meet the standard principles set out in the '2012 Games Impacts and Legacy Evaluation Framework' ('Legacy Evaluation Framework')

It was thought that the Meta-Evaluation therefore holds significant potential to advance methods more widely, particularly in terms of demonstrating how meta-evaluation can be employed practically in order to:

- Develop a framework for identifying, mining and aggregating data within a disparate body of existing evaluations;
- Inform better policy making and improve value for money; and
- Create a platform for more robust evaluation and research practice (in the field of mega events) in the future.

In response to this opportunity, the ESRC and the ECORYS Research Programme provided additional funding for a parallel research project to both help advance methods of meta-evaluation whilst improving the outcomes of the Meta-Evaluation itself.

Ecorys UK and Grant Thornton convened a team including four leading evaluation experts from the UK and the Netherlands with in-depth knowledge of evaluation methods, including meta-evaluation and meta-synthesis research, to develop a research specification and assist with conducting the research. These include:

[1] The ESRC is an independent UK non-departmental public body with an international reputation for supporting high quality research in social and economic issues, its commitment to training world-class social scientists and its role in disseminating knowledge and promoting public understanding of the social sciences.

- David Gough, Director of the Social Science Research Unit (and its EPPI-Centre) and Professor of Evidence-informed Policy and Practice at the Institute of Education, University of London;
- Steve Martin, Professor of Public Policy and Management at Cardiff Business School;
- Ray Pawson, Professor of Social Research Methodology in the School of Sociology and Social Policy, University of Leeds; and
- Henri de Groot, Professor to the Department of Spatial Economics and program coordinator of the BSc in Economics and Business, both at the Free University of Amsterdam, the Netherlands.

Jonathan France at Ecorys has managed the research project, working closely with Stephen Gifford and George Barrett, project leads of the Meta-Evaluation at Grant Thornton and Ecorys respectively, to ensure synergy with the wider study.

1.2 What is meta-evaluation?

The term 'meta-evaluation' was coined more than 40 years ago by Michael Scriven (1969). In simple terms, meta-evaluation means the 'evaluation of evaluations'.

A systematic literature search of peer-reviewed journals in 2009 identified just eighteen meta-evaluation studies, as well as some ambiguity about what 'meta-evaluation' actually involves (Cooksy and Caracelli 2009). For some, meta-evaluation refers to the study of the nature of evaluation. For others meta-evaluation is the setting of quality standards and applying these standards to interrogate the methodological integrity of evaluations, the process behind them, and the reliability of their findings. This can shed new light on good practice in the policy and practice of evaluations, while also raising questions about their limitations. The emphasis placed on processes and findings varies between studies. Some are primarily a quality assurance check on the approaches adopted by previous studies. However, meta-evaluation may also be interpreted as, or form the precursor to, the aggregation of data from existing evaluations. These meta-evaluations are concerned with bringing together the evidence from a range of studies and exploring implications for policy and practice and so overlap in purpose and methods with broad-based systematic mixed-methods reviews ('synthesis studies') and methods for testing the evidence for policy programmes (see Section 3 for a fuller discussion of these three types of meta-evaluation).

The starting point for this study is that meta-evaluation can be seen as a combination of evaluation science and methods of research synthesis. It involves consideration of the methods for identifying relevant primary research studies, methods for assessing their quality *and* relevance (Gough 2007), techniques for bringing together and interpreting empirical data collected by studies undertaken for different purposes and in different ways, and approaches to communicating with the audiences for meta-evaluation findings.

By considering both issues of quality and relevance, the weight of evidence that a study brings to the Meta-Evaluation of the Olympic and Paralympic Games can thus be assessed, prior to the synthesis of empirical results and aggregation of the overall impacts on beneficiary groups and other stakeholders.

1.3 Study Methodology

The research questions to be answered through the methods development study, agreed with ESRC, include:

- How can we better define and conceptualize meta-evaluation/analysis?
- What are the lessons from conducting previous meta-evaluations (at home and internationally) and how can meta-evaluation be improved?
- How can these lessons be applied to the Meta-Evaluation of the 2012 Olympic and Paralympic Games, in order to enhance methodology (and to help create an improved/exemplar model for measuring the impact of future mega-events)?

- What are the practical lessons from undertaking the Meta-Evaluation of the 2012 Olympic and Paralympic Games itself, which can advance methods of meta-evaluation?

The methodology to date has included:

Team briefing: the methods development study commenced with an in-depth briefing session for the research team to outline the main objectives, activities, challenges and opportunities in relation to the 2012 Games meta-evaluation, based upon the Project Initiation Document (PID) and key issues emerging from the scoping stage of the study. This ensured that the subsequent methods-development work for ESRC would be grounded in the context of the overall study, and that research team members were able to tailor the focus of their work towards the specific questions and issues facing the meta-evaluation team. The output of the meeting was a refined version of the research specification.

International literature review: a detailed review of the existing academic literature on meta-evaluation theory and practice was carried out in order to clarify definitions, outline processes of meta-evaluation (for systematic review and data synthesis), and to identify relevant studies and their lessons for the Meta-Evaluation of the 2012 Olympic and Paralympic Games. This review is included in sections two and three of this report.

Roundtable discussion on methods: two roundtable discussions were convened between the academics and operational members of the meta-evaluation team. The discussion groups examined the strengths and weaknesses of the approaches to meta-evaluation identified through the review, and how these might be applied to the 2012 Games meta-evaluation (and specifically to the early methodological scoping work and the development of logic models and theories of change). The outcomes of these discussions also informed the methods development study itself, through for example identifying specific questions to be put to the wider research community.

Consultation with the international research community: primary research with 13 experts drawn from the US, UK, and other European countries who have direct experience of conducting meta-evaluation and meta-analyses studies in order to assess in more detail the strengths and weaknesses of their studies and the practical lessons learnt, and to collate examples of useful research tools and frameworks. The analysis of these interviews is included in section four of this report.

Analysis and reporting: using the findings from the literature review, roundtable discussions and primary research, a set of recommendations and guidelines on the stages and steps involved in conducting meta-evaluation were developed. These focus on the methods and types of tools to be used by the Meta-Evaluation of the 2012 Olympic and Paralympic Games , in relation to the collation, review and synthesis of sources of evidence and the reporting of results (section five).

2 Literature on mega-events

Prior to the review of the literature on meta-evaluation, a number of reports of evaluations of previous Olympics and other large cultural and/or sporting events were examined. The objective was to understand the rationale, objectives and scope of such studies, as well as their some of their organising principles. The sample was therefore purposive and not exhaustive, and much of the material identified took the form of reports rather than peer reviewed papers.

The studies included attempt to bring together evidence from a variety of sources (including other evaluations) in order to provide an overview of the impacts of mega-events. Some provide a brief description of methods that have been employed by the studies they draw on but none of the studies undertake any detailed analysis of their strengths and weaknesses of the works they reference. They are therefore syntheses (the third type of meta-evaluation identified above). However, they do highlight some important methodological issues which are relevant to the Meta-Evaluation of the 2012 Olympic and Paralympic Games.

2.1 Objectives of mega-event evaluations

The studies reviewed illustrate the importance of being clear about the purpose (or intended outcomes) of mega-events because this in turn enables evaluators to develop criteria against which success can be assessed. This is not an easy task for four reasons:

- First, most mega-events have multiple objectives.
- Second, their stated objectives evolve over time.
- Third, different groups articulate different kinds of objectives.
- Fourth, outcomes may be negative as well as unanticipated.

The history of the modern Olympic Games illustrates this (Vigor *et al.* 2004). Three very different emphases have been to the fore at different times over the last 100 years:

Peace and understanding - De Coubertin's establishment of the Summer Games at the turn of the last century was motivated at least in part by a desire to counter rising nationalist tensions by bringing nations together in sports participation.

Economic impacts - By the 1980s and 1990s the Games had become highly commercialised. The Los Angeles and Atlanta Games are seen as prime examples of Games which serve a business sector agenda, but other host cities (notably Barcelona) used the Games as centrepieces for ambitious infrastructure projects and urban regeneration strategies.

Sustainability and legacy – From the Sydney Games onwards environmental sustainability became an important objective. London is also the first city selected to host the summer Games since changes in the IOC charter which mean that it now places much greater emphasis on the concept of longer-term 'legacy'. This makes the identification of appropriate legacy indicators a particularly important issue for the Meta-Evaluation of the Impacts and Legacy of the 2012 Olympic and Paralympic Games.

2.2 Multiple Legacies

There are though competing definitions of what constitutes a 'legacy', and different stakeholders will place the emphasis on different aspects (Shaffer *et al.* 2003). It may depend for example, on which political, commercial or community group is asking the question, and why. These issues will need to be taken into account in the meta-evaluation of the 2012 Games. Possible legacies may include for example:

- A debt free Games (emphasised in particular by the IOC)
- Accelerated regional development (an outcome of particular interest to the previous Labour Government and to the Greater London Authority)
- Promoting a positive image of London and sustaining the city's 'competitive edge' (an objective emphasised by the current Coalition Government and by the business community, particularly the conference, hospitality and events sector)
- Fixing London's transport infrastructure problems (a preoccupation of the media and a priority for many Londoners and commuters)
- Addressing employment and social problems in deprived communities (an important focus for boroughs and residents in the Lower Lea Valley).
- Boosting participation in sport and enhancing sports infrastructure (championed by both recent UK Governments, sports bodies such as Sport England, and sportsmen and women themselves).

The aspirations attached to different mega-events also reflect the contexts in which they are staged (Garcia *et al.* 2010). Issues of national identity are for example particularly poignant for countries that are emerging from difficult periods in their national history. The Barcelona Games were for example seen as important because they took place as Spain emerged from a period of dictatorship. Similarly, the Rugby World Cup was regarded as a defining moment in post-apartheid South Africa.

In recognition of their multiple objectives and scale, most previous evaluations of 'mega-events' have identified a range of different kinds of impacts and legacies. Almost all studies include:

- Economic;
- Social; and
- Environmental impacts.

Most recognise other types of impact or legacy as important, though they rarely agree on what these are. Indicators used in previous studies include:

- Improvements in governance capacity;
- Promoting national and/or regional identities;
- The development of employment and skills;
- Building up of social capital (for example through volunteering programmes);
- Place marketing, reputation management and branding; and
- Inclusion and well-being.

Studies typically analyse each key objective or legacy separately, frequently including a chapter on each major category of impact. However, within these chapters or themes multiple objectives or legacies will need to be pared down and each sub-set will on closer examination turn out to contain multiple ambitions which will also need to be sifted and prioritised.

2.3 Timescales

Some evaluations provide snap-shot assessments, but there is wide agreement in the literature that impacts and legacies really need to be evaluated over time (London Assembly 2007). There is also considerable scepticism about retrospective evaluations which rely on recall of events. The preferred methodology is therefore longitudinal analysis over a period of several years.

Some studies suggest that different kinds of impacts occur at different phases and that it is therefore useful to divide longitudinal studies into phases. The Olympic Games Global Impact approach identifies four:

- Conception;
- Organisation;
- Staging; and
- Closure.

The Rand Corporation (undated) suggests using three periods:

- Planning;
- Delivery; and
- Legacy.

It may be that different kinds of impact measures and meta-evaluation activity are needed at these different stages. For example during the planning phase evaluators are likely to focus on activities such as agreeing on the Games' objectives, agreeing assessment criteria, developing theories of change, constructing baselines, identifying relevant sources of evidence about impacts (and potential gaps in the data), working with other evaluators to make sure the data they need will be gathered, and conducting a formative assessment of impact. During the implementation phase they may be engaged in data gathering to help assess the short-term and immediate impacts of staging the event, whilst working with other evaluators to help ensure that their methods are robust, and potentially in conducting additional primary research. During the legacy phase they may gather further data and assess and pull together the available evidence to provide an ex post impact assessment.

2.4 Breadth of analysis

Many studies differentiate between direct and indirect impacts, particularly in respect of economic effects. Many suggest that indirect impacts are much more difficult to measure and therefore that casting the evaluation net too wide (for example using formulae to estimate second and third order multiplier effects) is likely to reduce the rigour of a study. Clearly there is a difficult trade-off to be made. To take too broad and too long a view would risk undermining the reliability and credibility of any meta-evaluation. But to focus too narrowly would be to miss many of its anticipated benefits which are by nature indirect and possibly even intangible (Langen and Garcia 2009).

There is also a sense from the literature that mega-events often leave some sort of overall lasting 'impression'. But this is difficult to pin down and it is clear that some of the factors which contribute to it can not be managed by host cities and countries. Drug scandals, terrorist acts or even the prevailing weather conditions may be put down to (good or bad) 'luck'. However, it may be legitimate for evaluations to explore the extent to which such potential impacts were anticipated, planned for (through design of quality assurance and resilience mechanisms) and reacted to when they occurred. The unintended impacts and consequences of mega-events are therefore frequently also a focus of such studies.

2.5 Distributional effects

Previous studies highlight issues of who pays for and who benefits from mega-events. This includes issues of which social groups benefit and the impact on localities of hosting events. In the short term issues such as who gains jobs in the construction phase loom large. In the longer term there are questions about whether local people benefit from improvements in infrastructure and the provision of new stadia and other sport facilities. In theory Londoners should benefit from a range of physical legacies but in the past in some cities, escalating property values associated with urban renewal resulting from or accelerated by a mega-event have driven locals out of the area (Smith 2008).

Some studies have emphasised the importance of including locals' views in evaluations of mega-events, and some have experimented with methods which assess the public's willingness to pay for events as a means of testing the perceived value which the public places upon them.

2.6 Integrating evaluative frameworks

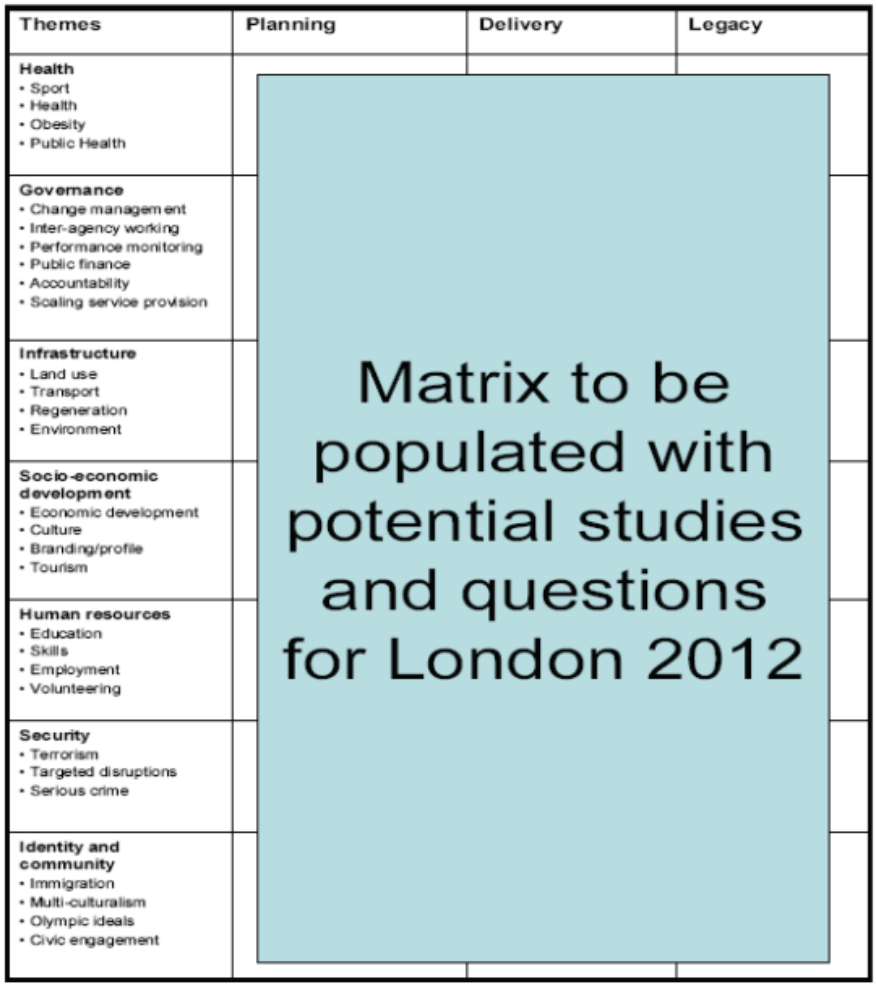
Different kinds of mega-event impacts and legacies require different measures and possibly evaluation methodologies, so it is challenging to find a grand conceptual amalgam capable of reflecting all ambitions.

The literature nonetheless offers some possible pointers to frameworks that might help to structure the Meta-Evaluation of the 2012 Olympic and Paralympic Games. Rand Europe (undated) suggests commencing with a matrix with key 'themes' (in essence potential 'families of impact') identified on one axis and the three phases of mega events listed on the other axis (see Figure 1 below). They argue that this can then be used to help define evaluation questions and to build alternative outcome scenarios.

However, it is also clear that mega-event evaluations need to consider the interactions - mutual contributions and/or contradictions – between these different themes. This implies that the logic models developed through the evaluation process should be used to identify how these high-level objectives and outcomes are inter-related.

More generally, the literature on broad based mixed-methods and theory-driven systematic reviews provides a model for how the data can be interrogated to address questions of the outcomes of mega-events, as discussed in the following chapter.

Figure 2-1: Evaluation matrix for mega-events



3 Literature on meta-evaluation

3.1 Definitions of meta-evaluation

The word evaluation refers to judgments of something's value, quality, importance, extent, or condition (Encarta dictionary), though it is also often used to refer to research evaluating whether some service or programme has achieved its objectives and not achieved some undesired outcomes (see Scriven 1999 on fields of evaluation).

The word 'meta' has many meanings and often means about or beyond (Thomas 1984). The term 'meta-evaluation' was coined more than 40 years ago by Michael Scriven who offered the straightforward definition of this activity as "the evaluation of evaluations" (1969). As has already been mentioned in section 1, this can mean at least three different types of evaluation depending on how evaluations are being evaluated.

(i) The meta-theory of evaluation

Scriven (1969) states that one type of meta-evaluation is 'the methodological assessment of the role of evaluation'. In other words, this is the evaluation of the nature and purpose of evaluation. The pursuit of any science raises questions about its foundations and first principles. Under this meaning, meta-evaluation raises questions about meta-theory (basic logic, strategy, methodology, epistemology, ontology of evaluation) on issues such as: the prime function of evaluation; what can and cannot be evaluated; how (un)certain is the evidence; the extent that findings are transferable; and how we should understand causation in policy analysis. Such meta-theory is fundamental both to the meaning of evaluation but also to the two other main forms of evaluation.

(ii) Meta-evaluation of the quality and standards of evaluation studies

Scriven (1969) argues that a main form of meta-evaluation is 'the evaluation of specific evaluative performances'. In other words, this is the evaluation of the quality of evaluation studies. This can be a concern for the usefulness of a study, the adequacy of the research team or organization, or the assessment of the strengths and weaknesses of a method and the creation of methodological standards for evaluation. It can take both formative and summative forms. One definition of these forms of meta-evaluation is:

"Meta-evaluation is the process of delineating, obtaining, and applying descriptive information and judgmental information about an evaluation's utility, feasibility, propriety, and accuracy and its systematic nature, competence, integrity/honesty, respectfulness, and social responsibility to guide the evaluation and publicly report its strengths and weaknesses. Formative meta-evaluations— employed in undertaking and conducting evaluations—assist evaluators to plan, conduct, improve, interpret, and report their evaluation studies. Summative meta-evaluations — conducted following an evaluation — help audiences see an evaluation's strengths and weaknesses, and judge its merit and worth." (Stufflebeam 2001 p183)

(iii) Meta-evaluation synthesis of findings of evaluations

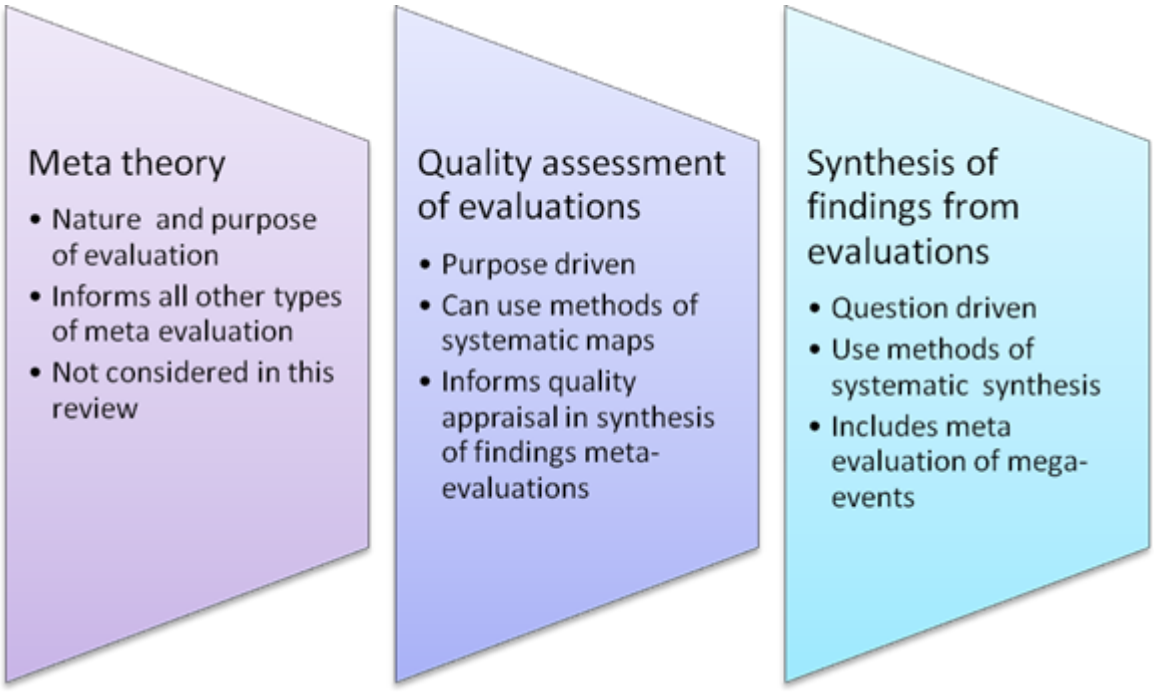
Another type of meta-evaluation is the synthesis of the findings of individual studies to answer an evaluation research question. In other words, this is the combination (or aggregation) of multiple evaluation studies. Evaluation is often of an individual occurrence of a single intervention. In meta-evaluation, there is an opportunity for the evaluation of multiple evaluations and so the unit of analysis becomes larger segments of policy making. The logic is that modern social and behavioural interventions have a history. They are tried and tried again and researched and researched again, and it therefore makes sense to try to identify common themes and lessons from this collective experience.

This process often includes interrogation of the methodological integrity and the reliability of the findings of the individual studies and so is informed by quality standards of evaluation (as in the quality standards definition above).

All three forms of meta-evaluation have value. Meta theory raises fundamental issues about the nature and purpose of evaluations and is the building block for evaluation science. Evaluations of quality standards develops good practice in both methodological and policy terms and can raise important questions about the limitations of methods and enables policy makers and others to determine whether to take notice of the findings of evaluations. The synthesis of multiple evaluations results in a fuller understanding of the effectiveness of a policy initiative.

Given the task in hand, to support the Meta-Evaluation of the 2012 Olympic and Paralympic Games, and to derive learning from the process, this review is not concerned with the broader meaning of evaluating evaluation science and the development of a meta theory of evaluation. It is concerned with the other two forms of meta-evaluation: quality and standards of evaluation and the synthesis of existing evaluations.

Figure 3-1: Three Main Types of Meta-Evaluation



3.2 The literature on meta-evaluation

The aim of this report has been to identify some key messages from the literature, in order to inform the development of the methodology for the Meta-Evaluation of the 2012 Olympic and Paralympic Games. It is not an exhaustive search of the literature but a purposive search and configuring of variation in forms of meta-evaluation. The literature for this review was identified from two sources:

- First, a systematic search was made of bibliographic databases for papers that included the terms 'meta-evaluation' or 'metaevaluation' or 'meta evaluation'. The databases were from the British Humanities Index, Medline, Social Science databases and Web of Science. The search identified 204 potential papers including duplications.
- Second, 14 papers were identified from a course on meta-evaluation at Western Michigan University.

The literature included methodological papers discussing the definition of meta-evaluation and papers reporting the results of meta-evaluations. It also included reports and papers which did not describe themselves as 'meta-evaluation' but had nonetheless analysed the often complex and inter-related impacts of 'mega-events'.

The search of the literature found examples of both quality standards and the synthesis of prior evaluations. Both these forms of meta-evaluation can use methods of systematic reviews. The broader literature on systematic reviews (including statistical meta-analysis of findings of studies of the impact of interventions) is very large and was not searched for during this review, though the authors are aware of and refer to some of this literature in this report.

3.3 'Quality of methods' meta-evaluations

This form of meta-evaluation develops standards for methods of evaluation, applies these to inform the planning of evaluations and in assessing the quality of evaluations, and further develops standards. Such 'evaluation of specific evaluative performances' can take several forms depending on the aims. The general approach is that meta-evaluation can be done using the same logic and sometimes methods used in primary evaluation (Shadish 1988).

(i) Aims and methods of 'quality of methods' meta-evaluations

There are many reasons why one might want to evaluate the methods of an evaluation. It may be to assess the trustworthiness of the study, to audit and develop methods of evaluation (and inform future research plans) or to develop quality standards of evaluation.

Stufflebeam suggests the following steps for carrying out quality assessment meta-evaluations in practice. (2001, p191):

Structure for Identifying Alternative Meta-evaluation Procedures

- Determine and arrange to interact with the meta-evaluation's stakeholders.
- Staff the meta-evaluation team with one or more qualified evaluators.
- Define the meta-evaluation questions.
- Agree on standards, principles, and/or criteria to judge the evaluation system or evaluation.
- Develop the memorandum of agreement or contract to govern the meta-evaluation.
- Collect and review pertinent available information.
- Collect new information as needed, including, for example, through on-site interviews, observations, and surveys.
- Analyze the qualitative and quantitative information.
- Judge the evaluation's adherence to appropriate standards, principles, and/or criteria.
- Convey the meta-evaluation findings through reports, correspondence, oral presentations, etc.
- As needed and feasible, help the client and other stakeholders to interpret and apply findings.

(ii) Trustworthiness of study findings

This is the assessment of the usefulness of a study to determine whether the results of a study can be relied upon. An example would be the refereeing of an article reporting an evaluation submitted to a journal for publication. The referee process managed by the journal editors would assess the worth of the study for publication. Another example would be the appraisal of the worth of a study for inclusion in a synthesis of many studies. In this way, the quality standards form of meta-evaluation is used in the synthesis of studies form of meta-evaluations.

(iii) Audit and development of methods

This involves the assessment of the adequacy or audit of a series of studies usually by a research team or organization, for a specific purpose (Green et al 1992, Schwandt 1992, Schwarz & Mayne 2005). An example would be a funder deciding whether the previous evaluations by an organization were of sufficient quality to persuade them to provide further research funding. Another example would be an organization making a study of the process of evaluation in its work (for example, Bornmann et al 2006, 2010). A further example, would be an organization reviewing its own research to decide on further plans such as further methods capacity development or future research plans (as in the boxed example on Cooksy and Caracelli). Organisations might also seek to develop a template for evaluation studies, which they commission to ensure that their evaluations are helpful to policy formation (for example Department for International Development in 2008, which reviewed the evaluation methodology used in its 'country studies' and sought to strengthen the methodology, using experience from comparable evaluations in other parts of Whitehall and internationally).

Box 3-1: Consultative Group on International Agricultural Research (CGIAR)

Aims: CGIAR assessed the evaluations of member organisations in order to ask: (i) What is the substantive focus (e.g. type and level of impact examined) of the studies conducted by the CGIAR centers?; (ii) What is the methodological quality of the studies?; (iii) Are there enough studies of high enough quality to support a synthesis across studies?

Method: (i) All 87 evaluation reports were coded for the substantive and methodological characteristics of each study; (ii) Assessment of each study's credibility by comparing information about its methodological characteristics to the inferences that were drawn about programme impact; (iii) An analysis of the reasons that were documented for positive or negative assessments of credibility.

Results: (i) Large variety in the focus and methods of studies; (ii) Lack of transparency of reporting meant quality could not be clearly assessed; (iii) Not possible to synthesize such heterogeneous studies of unknown quality.

(Cooksy and Caracelli 2005)

The review of the methods within a programme of work undertaken systematically, such as of CGIAR above, is a form of **systematic map review**. Studies are only included if they meet the inclusion criteria and they are then coded[2] in order to obtain an overview of the studies.

This approach has been taken a step further with the assessment of the methodological aspects of a specific field of study using data from multiple systematic reviews; i.e. an analysis of the coding of studies across a series of systematic reviews. If each review contains many studies then the total sample of studies included can be very large. This approach has been used to assess the methods of randomized control trials and their effects on statistical meta-analysis (synthesis) and is called meta-epidemiology (as in the boxed example on Oliver et al 2010).

[2] The process of combing data for themes, ideas and categories and marking similar data with a code label in order that they may be easily retrieved at a later stage for comparison and analysis.

Box 3-2: Randomised controlled trials for policy interventions

Aims: To assess whether randomized and non randomized studies of similar policy interventions have the same effect size and variance.

Method: Investigating associations between randomization and effect size in studies coded for systematic reviews (meta-epidemiology)

Results: Non randomized trial may lead to different effect sizes but the effects are unpredictable.

(Oliver et al 2010)

This approach has also been taken further in meta-evaluations that analyse the role that evaluation can play in influencing public policy. Bustelo (2003a), for example, assessed the role of evaluation processes in Spanish regional and national gender equality plans (see boxed example).

Box 3-3: Evaluation of gender mainstreaming

Aims: To analyse the evaluation processes of 11 public gender equality policies implemented between 1995 and 1999 in Spain.

Method: Evaluation processes evaluated against 6 criteria

Results: Ten main conclusions of: (i) lack of clarity in the evaluation purposes: were the evaluations of gender equality policies and the plans of action, or were they evaluations of women's status?; (ii) lack of a global vision of the public action taken for promoting gender equality: were the evaluations of the policies or simply of specific plans of action?; (iii) lack of recognition that evaluations are themselves political acts; (iv) the perception of evaluation as a secondary function: the important role women's agencies should play around policy evaluation; (v) the need to know exactly WHAT we want to evaluate: the "dictatorship" of the methodology and the techniques; (vi) importance of the institutional and co-ordination structures for evaluation; (vii) importance of timeliness; (viii) a clear deficit of "practical elaboration"; (ix) poor communication and dissemination processes; (x) a need for a greater resource investment in evaluation.

(Bustelo 2003)

(iv) Development of quality standards

This is the assessment of the strengths and weaknesses of a method in order to support the creation of new methods for evaluation and the professionalization of evaluation (Bickman 1997, Bollen et al 2005). This is a core academic activity with numerous academic journals concerned with testing and development of methods from different research paradigms. This has led some to develop quality standards for evaluation such as those developed in the United States by the Joint Committee on Standards for Educational Evaluation[3] (as in the boxed example on Yarbrough 2011) and the Evaluation Centre at Western Michigan University[4] plus many others in the United Kingdom[5] and further internationally.

[3] http://www.jcsee.org/
[4] http://www.wmich.edu/evalctr/checklists/
[5] http://www.evaluation.org.uk/resources/guidelines.aspx

Box 3-4: Standards for Educational Evaluation

Aims: Joint Committee on Standards for Educational Evaluation develops standards for educational evaluations to promote evaluations of high quality based on sound evaluation practices and procedures

Method: Needs assessments, reviews of existing scholarship, involvement of many stakeholders, field trials, and national hearings.

Results: Thirty standards within five main categories of: i) Utility (evaluation processes and products valuable in meeting their needs); ii) Feasibility (effectiveness and efficiency); iii) Propriety (proper, fair, legal, right and just in evaluations); iv) Accuracy (the dependability and truthfulness of evaluation representations, propositions, and findings, especially those that support interpretations and judgments about quality); and v) Accountability (adequate documentation of evaluations and a meta evaluative perspective focused on improvement and accountability for evaluation processes and products).

(Yarbrough et al 2011)

This area of work develops new methods, develops standards and capacity to use and report such methods, and can also be used to critically appraise the quality of individual or multiple studies.

(v) Dimensions of difference in 'quality of methods' meta-evaluations

There are many other ways in which the meta-evaluation of the quality of research methods can vary.

A major source of variation is the basis for the evaluation of methods. This may be driven by a number of different epistemological positions and by very different purposes. The evaluation may not, for example, be simply based upon quantitative paradigms with pre-specified criteria of value but may also be based on more emergent qualitative criteria (for example, Curran et al 2003, Maxwell 1984). Similarly, there can be variation within a meta-evaluation; the aims and research position taken by the evaluation may or may not be in line with the aims or assumptions of the researchers whose research is being evaluated (see also section on quality appraisal).

In a recent survey of eighteen meta-evaluations of single studies Cooksy and Caracelli (2009) found that 5 were assessed according to quality standards, 3 using criteria developed specifically for that meta-evaluation, 3 used the criterion of trustworthiness based on the confirmability and dependability of the findings, and 7 used inductive approaches of emergent criteria for quality related to the extent to which the evaluation addressed the purposes of the programmes.

Whatever the overall aims of a quality of methods meta-evaluation, it can also differ in the phase of the research process that it focuses upon. It can focus on the planned methods of evaluation (**design meta-evaluation**), on how these plans were implemented in practice (**process meta-evaluation**) or on the results of the evaluation (**results meta-evaluation**) (Bustelo 2003b), or all three. This will of course affect the criteria used to make the evaluative assessments. A related area of variation is the role of the evaluator. They may be the researchers or their colleagues and part of an internal appraisal. Alternatively, they may be external to and independent from the primary evaluations.

Another type of variation is the timing of the meta-evaluation. It may occur before, during and/or after the completion of the study being considered. It may be formative and undertaken whilst the study is planned or underway. This might include feedback during the process of the evaluation of a planned or ongoing study to improve the manner in which the evaluation is being conducted (Stufflebeam 1981, Hanssen et al 2008). Alternatively, the evaluation of the study may be summative and undertaken once the study is complete. The quality analysis of the studies may involve an analysis of the raw data in the studies or replications of studies. Some of these choices are listed in the table below (re-ordered table from Cook and Gruder 1978, p 17).

Simultaneous with primary evaluation	Data not manipulated	Single or multiple data sets	Consultant meta-evaluation
	Data manipulated	Single data set	Simultaneous secondary evaluation of raw data
		Multiple data sets	Multiple independent replications
Subsequent to primary evaluation	Data not manipulated	Single data set	Essay review of an evaluation report
		Multiple data sets	Review of the literature about a specific programme
	Data manipulated	Single data set	Empirical re-evaluation of an evaluation or programme
		Multiple data sets	Empirical re-evaluation of multiple data sets about the same programme

(vi) Summary of differences

The differences in 'quality of methods' meta-evaluations can be summarized as follows:

Box 3-5:

Aims. These may relate to: (i) evaluating the quality of a study to determine its trustworthiness; (ii) a broader remit of auditing the quality of studies and enabling the development of their quality; or (iii) the development of quality standards to make such trustworthiness and audit assessments

Evaluation phase. Meta-evaluation may focus on: (i) the design of a study; (ii) the process by which a study is undertaken; or (iii) the results of an evaluation study

Criteria: The criteria are the bases on which the evaluation judgments are made (such as quality standards)

Independence of evaluator. The meta-evaluator may be: (i) external and independent; or (ii) internal and related to the evaluation being evaluated

Timing. Meta-evaluation may be: (i) concurrent and formative; or (ii) after the evaluation and summative

Manipulation of data. The data may be: (i) used as reported by the evaluations; or (ii) re-analysed

Methods. A range of procedures may be used to undertake the meta-evaluation (these methods are covered in more detail in chapter 5).

Although there are many types of 'quality of methods' meta-evaluation, it is possible for one particular study to combine aspects of these different types. Also, it is possible for the meta-evaluation to reflect on its own methods and thus be a 'meta' meta-evaluation of the quality of methods, as in the study by Madzivhandila et al (2010, see boxed example).

Box 3-6: Meta-evaluations in government and government institutions

Aims: To review: (i) the quality of the impact assessment evaluations of the Australian Centre for International Agricultural Research (ACIAR); and (ii) the process of reviewing methods and quality assessment.

Method: Retrospective and real time evaluations of the ACIAR evaluations using Program Evaluation Standards.

Results: there was non-use or low use of some standards in the 19 evaluation studies: evaluation stakeholders identification; practical procedures; political viability; formal agreements; rights of human subjects; human interactions; fiscal responsibility; analysis of qualitative information; and the use of meta-evaluation. The lessons learned from the meta-evaluation are used to develop proposed further systematic meta-evaluations.

Madzivhandila et al (2010)

3.4 Synthesis meta-evaluations

This form of meta-evaluation synthesizes the findings of multiple evaluations to undertake one large evaluation. If this is done systematically then this form of meta-evaluation is a form (or many forms of) systematic review. Systematic reviews bring together existing research studies focused on a specific question or intervention to better understand what we know from that literature. This is a form of secondary research and requires specification of the research questions (and its assumptions), and explicit rigorous methods of identification, appraisal, and synthesis of study findings to answer the review question (Gough and Thomas 2012).

(i) Aims and methods of synthesis meta-evaluations

The particular type of synthesis meta-evaluation will depend upon the approach to evaluation and the specific evaluation question being asked.

(ii) Aggregating and configuring reviews

The challenge has been taken up in slightly different ways and it is useful as a starting point to distinguish between two approaches.

First are systematic reviews (or research synthesis or, confusingly, meta-analysis) that starts from the premise that broadly the same intervention has been tried many times in different locations. Evidence from previous research on all/many such instances is uncovered. Then, using a variety of different methods of summing or synthesising the evidence, the review will attempt to assess the efficacy of that family of programmes. The emphasis is on precision of measuring efficacy usually through attempting homogeneity of interventions and measures and effect. These reviews are essentially combining (aggregating) the findings of individual studies and their measurements to create an overall summary finding across studies (Voils et al 2008, Sandelowski et al 2012). They can also examine how to arrange and understand (configure) variation in effects within the studies using techniques such as meta-regression.

Second are systematic reviews that seek to take account of the complexity and contingent nature of interventions. Interventions are seen as being strongly influenced by their political, policy, cultural and social settings. Hence meta-evaluations focus on the evolution of programmes, interactions among them, and/or the effects of the wider environments in which they are enacted, and are often concerned with questions about the collective fate of interventions. The emphasis is on the heterogeneity of interventions and effects and the consequences of this for the generalisability of review findings. The reviews are essentially configuring findings to understand empirical and conceptual patterns (Voils et al 2008. Sandelowski et al 2012).

This simple binary division helps to distinguish the main types of review, though in practice specific review types may contain degrees of both types of review and thus different synthesis methods.

(iii) Experimental assessment of the efficacy of an intervention

This includes systematic reviews that aggregate results of quantitative experimentally controlled impact studies to test theories of impact (or 'what works?'). If statistical data is available for synthesis then these reviews are called statistical meta-analyses (or just meta-analysis for short) (see boxed example on Petrosino et al 2002). They may also employ statistical methods such as meta-regression to examine internal variation between the results related to variation in intervention, participants or context. In other cases there may only be correlational data or no statistical data available and synthesis is based on grouping textual data. All of these reviews tend to be testing theories using pre-specified concepts and methods.

Box 3-7: Scared straight

Aims: To assess the effects of programmes comprising organised visits to prisons by delinquents and children in trouble aimed at deterring them from criminal activity.

Method: Statistical meta-analysis.

Results: The analysis shows that the intervention appears to be more harmful than doing nothing.

(Petrosino et al 2002)

(iv) Testing of causal theories and realist synthesis

Experimental evaluation of efficacy can be based on a detailed theory of change (causal effects) or may simply be testing whether a difference is found with no theory as to why this might be so (a 'black box' approach). Theory testing approaches are more concerned with hypothesizing and testing and then refining theories of what mechanisms explain why interventions work (i.e. have the outcomes been delivered as intended), and in what contexts. These may be relatively simple theories or may be more complex and their study may involve an ongoing sequence of studies and multi component reviews.

Realist synthesis is one particular form of theory testing review that unpacks and arranges (configures) the theoretical and practical components of the theory/policy being evaluated and then uses iterative methods to explore data to test these theories, based upon gathering together existing evidence of success. A theory or policy initiative may be successful in some circumstances and not others and realist synthesis examines the logic models that underlie these variations in practice (Pawson, 2006 and see boxed example on Pawson 2002).

Box 3-8: Megan's Law

Aims: To assess whether the US sex offender notification and registration programme works.

Method: Realist Synthesis

Results: Megan's Law is a programme with a long implementation chain iterative in its impact. The complexity of decision-making compounds at every point with the result that there is little guarantee of uniformity between cases as they proceed through the registration and notification process. Offenders with identical records may have very different experiences. The programme thus achieves some of its objectives in some cases but in many cases does not.

(Pawson 2002)

(v) Conceptualizing experience, meaning and process: qualitative synthesis

Efficacy reviews tend to use aggregative theory-testing methods. Other reviews configure results of empirical or conceptual studies to generate or explore theories about *experience, meaning and process*. Examples would be reviews of research on the processes by which things work, and these may include qualitative research and conceptual data and thus non-statistical and more qualitative forms of synthesis (Rodgers et al 2009). Such reviews also interpret, organise and configure concepts using iterative methods of review rather than using pre-specified concepts and methods. There are often many very different theories relevant to the study of a social issue and so a configuring review may assist in analysing the theoretical landscape before testing any

individual or group of theories (or developing new theories to test) (Gough et al 2012). Importantly, these different types of review can be combined; even if an aggregative theory testing review is being undertaken, it may be helpful to have additional data to interpret and understand the meaning of the data.

One example of such an approach is **meta-ethnography** where the reviewer is akin to an ethnographer undertaking primary research. However, instead of experiencing real world situations directly, the data for the reviewer are previous ethnographies (and other types of in-depth qualitative study). This involves examining the key concepts within and across studies through a process called reciprocal translation, which is analogous to the method of constant comparison used in primary qualitative data analysis[6]. This process creates new interpretative constructions and a line of argument to create higher order 'meta' ethnographic interpretations that could not be achieved by the individual primary studies alone (Noblitt and Hare 1988) (see boxed example on Britten et al 2002).

Box 3-9: Resistance to taking medicines

Aims: To assess how the perceived *meanings* of medicines affect patients' medicine-taking behaviour and communication with health professionals.

Method: Meta ethnography

Results: These include third order interpretations that include but go beyond the findings in individual primary studies: (i) Self-regulation includes the use of alternative coping strategies; (ii) Self-regulation flourishes if sanctions are not severe; (iii) Alternative coping strategies are not seen by patients as medically legitimate; (iv) Fear of sanctions and guilt produce selective disclosure.

(Britten et al 2002)

Some configuring reviews exploring and generating theory take a more critical stance to theory. **Critical interpretative synthesis** (Dixon Woods et al 2006) is similar to meta-ethnography in applying principles of qualitative enquiry (particularly grounded theory[7]) to reviewing and developing a conceptual argument through the process of the review. However, it takes a more critical interpretative approach to the epistemological and normative assumptions of the literature that it reviews. The reviewers' 'voice' in problematizing and interpreting the literature is stronger than in meta-ethnography.

Another critical approach to configuring conceptual reviews is **meta-narrative reviews** (Greenhalgh et al 2005, see boxed example). The units of analysis in these reviews are the unfolding 'storylines' or narratives of different approaches to studying an issue over time; that is the historical development of concepts, theory and methods in each research tradition. These different narratives from different research approaches are first separated and mapped out and then brought together to build up a rich picture of the area of study. There are similarities to some aspects of meta-ethnography and critical interpretative synthesis in that different concepts are identified and then reinterpreted into a new argument.

[6] The process of returning to previously analysed text during coding of qualitative data, to ensure consistency of approach, and to identify new dimensions or phenomena.
[7] The generation of theory from data, rather than beginning with a hypothesis to be tested.

Box 3-10: Diffusion of innovations in health service organisations

Aims: To review the literature on how to spread and sustain innovations in health service delivery and organisation

Method: Meta narrative review

Results: A unifying conceptual model with determinants of innovation, dissemination, diffusion, system antecedents, system readiness, adoption/assimilation, implementation and consequences.

(Greenhalgh et al 2005)

(vi) 3.4.1.5 Mixed methods systematic reviews

Another strategy for reviewing complex issues is to undertake mixed methods reviews. These can mix methods within one review process (as does Realist Synthesis) or can separately review sub-questions and then integrate these together to provide an overall review.

Box 3-11: Barriers and facilities of healthy eating

Aims: To review what is known about the barriers to and facilitators of healthy eating amongst children aged four to 10 years old.
Method: Nineteen outcome evaluations were entered into a statistical meta-analysis, and the findings from eight studies of children's views were analysed through a thematic synthesis. The findings of both syntheses were then brought together to see whether interventions which matched children's views were more effective than those that did not.
Results: The sub-review on efficacy found a statistically significant, positive effect from health promotion. The sub-review on children's views suggested that interventions should treat fruit and vegetables in different ways, and should not focus on health warnings. Interventions which were in line with these suggestions tended to be more effective than those that were not.

(Thomas et al 2004)

(vii) Reviews of reviews

Another review strategy is to use previous reviews rather than primary studies as the data for the review. The resultant review of reviews may be of similar or different types of review and similar or different types of studies included in each review, which raises issues of mixed methods and heterogeneity in reviews (see boxed example on Caird et al 2010).

Box 3-12: The socioeconomic value of nursing and midwifery

Aims: To review what socioeconomic benefits can be attributed to nursing and midwifery with respect to: mental health nursing; long-term conditions; and role substitution.

Method: Thirty-two systematic reviews were available for inclusion within the review. The findings from reviews with similar topics were grouped and synthesised using a meta-narrative approach using where possible, review authors' pooling of data. Often, authors had presented findings in a narrative form and so the review of reviews' syntheses are themselves narrative in form.

Results: There was evidence of the benefits of nursing and midwifery for a range of outcomes. This was accompanied by no evidence of difference for other outcomes (statistical tests failed to demonstrate a significant difference between nurse/midwife-delivered interventions and those provided by others). An important finding of this review was that nursing and midwifery care when compared with other types of care was not shown to produce adverse outcomes. The included reviews rarely provided cost or cost-effectiveness data.

(Caird et al 2010)

(viii) Non systematic reviews of broad research questions

Some reviews aggregate and configure statistical or other forms of research data to address broadly-based questions and/or complex questions using some of the insights of systematic review but often without a specific review methodology. This may be because of resource

constraints in reviewing such broad questions and research material though some of these reviews do manage to follow systematic principles (for example, Ashworth et al 2004).

Such approaches are common in reviews of policy agendas and stratagems with broad aims or huge targets or grand philosophies (as in boxed example on Warwick et al 2009). Such policies are delivered via a range of different interventions and service modifications. Meta-evaluation enters here with the task of researching the collective endeavour. For example, evaluating healthy school initiatives or methods to reduce the population who are not in work or employment, or increasing voluntarism in the big society.

Box 3-13: Healthy Schools

Aims: To provide an overview of existing evidence on the effectiveness of healthy schools approaches to promoting health and well-being among children and young people

Method: An analysis of the research literature into major themes and findings.

Results: Successful programmes share a focus on: promoting mental health rather than preventing mental illness; securing long-term rather than short-term goals;

improving the whole school 'climate'; providing a wide range of opportunities for practising new skills; engaging with multiple sites including the school, the family and the community; delivering both universal and targeted activities .

(Warwick et al 2009)

Broad based reviews may also undertake the primary research that they analyse. They study a range of different services or organizations and then draw the findings together to provide a 'meta' overview. In some cases this might include an evaluation of the evaluation processes in the services or organizations being studied. In such cases, the study includes both major types of meta-evaluation; an evaluation of evaluation processes and a synthesis of these and other findings across the services/organization as in the Eureval (2008) study.

Box 3-14: Meta study on decentralised agencies

Aims: To increase the transparency of European agencies and the responsiveness to information needs of European institutions.

Method: (i) Evaluation of documents from and interviews with individual agencies on relevance, coherence, effectiveness and internal efficiency of the agencies, plus coherence of the evaluation requirements and practices; (ii) synthesis of the findings across agencies.

Results: Detailed results lead to conclusions on: relevance to needs; priority-setting; rationale; coherence with the EU policy served and coordination with the parent DG; coherence and coordination between agencies; coherence with non-EU bodies; effectiveness; cost-effectiveness; community added value; proximity and visibility; productivity; strategy-making; management methods; coverage of evaluation issues; needs of evaluation users; and use of evaluation findings and conclusions

(Eureval 2008)

Policy coordination of 'joined-up policy-making' is also a major aspiration of modern government. Researching the coordination (or otherwise) of the agencies who deliver an intervention is thus another meta-evaluative task (for example, the coordination of police, local authorities, youth and community services in the delivery of Anti Social Behaviour Orders). There is also policy sequencing, the optimal timing and sequencing of interventions. For example, smoking bans have been enacted on public transport, followed by office and indoor workplace restrictions, followed by smoke-free restaurants and finally bars, pubs, and gambling venues. Is public opinion thus primed for the next location - private cars?

(ix) Dimensions of difference in 'synthesis' meta-evaluations

The summary and examples provided above of several types of review that could be considered forms of meta-evaluation do not fully reveal the extent of variation that exists between different systematic reviews. This section describes some of these dimensions of difference (for more details see Gough and Thomas 2012; Gough et al 2012).

In general, reviews reflect the variation in approaches and methods found in primary research. They vary in their research paradigm and their underlying epistemology. The aggregative reviews of efficacy and Realist Synthesis both assume a realist epistemology where knowledge can approximate an external reality. Configurative reviews of conceptual data, however, may take an idealist stance where such an agreed external reality is not assumed (Barnett-Page and Thomas 2009). As already discussed, aggregative reviews tend to be theory testing, use pre-specified concepts and methods and seek homogeneity of data. Configuring reviews tend to generate or explore theory using iterative concepts and methods and seek heterogeneity.

Reviews also vary in their structure, whether they are simply a map of research or also a synthesis of findings from that map (or sub-map). Reviews can contain sub-reviews (as in the mixed methods reviews discussed above) or can be meta-reviews such as reviews of reviews (and also meta-epidemiology as discussed in 'quality of methods' forms of meta-evaluation).

Reviews can be of very broad or narrow questions and can be undertaken in great depth of detail or in a relatively less detailed way. The broader the question and the deeper the detail, the greater the challenge to manage the diversity of issues (as is often the case in the meta-evaluation of mega-events). Pressures of time and funding lead some to undertake rapid reviews which often need to be narrow and lacking in detail to be undertaken systematically with such little resource, or to lack rigor in method.

(x) Summary of differences

The differences in synthesis meta-evaluations can be summarized as follows:

- **Broad review type and methods**: aggregative reviews, which test the efficacy of interventions (methods of meta-analysis) or groups of interventions and their logic and contexts (methods of realist synthesis) against pre-defined theories and methods; and conceptualizing/configuring reviews, which tend to generate or explore theory, incorporating in particular methods of qualitative synthesis (as in the more iterative elements of realist synthesis, and meta ethnography, critical interpretive synthesis, and meta narrative reviews).
- **Research paradigm**: realist epistemology (not questioning that there is some form of shared reality to be studied) vs. idealist stance (not assuming that there is any reality independent of our experience)
- **Meta-reviews**: reviews combining reviews as in, for example, mixed methods systematic reviews (or sub-reviews), and reviews of reviews.
- **Rigour of review methods**: the relative degree of rigour that distinguishes a systematic review from a non systematic review; for example, a non systematic scoping of studies to inform a more systematic review (or a lack of rigour simply due to resource constraints).
- **Level of detail**: both systematic and non-systematic reviews can be narrow or broad in their focus, and more or less detailed, depending upon a combination of aims, available resources and the requirement for systematic methods. Where resources are limited, there may be a trade off between breadth and rigour of methods.

3.5 Conclusions from the literature review

The literature shows that meta-evaluations can vary widely in their purpose and methods and confirms the conclusion by Cooksy and Caracelli (2009) that the evaluation field does not have a common understanding of meta-evaluation practice.

The review of the literature revealed three main types of meta-evaluation: meta-theory; quality assessment of evaluations; and synthesis of findings of evaluations. It is the third of these, **synthesis of findings of evaluations**, that most clearly fits with the meta-evaluation of mega-events (and in turn the Meta-Evaluation of the 2012 Olympic and Paralympic Games). The synthesis of findings through systematic review often includes studies of separate examples of an event or situation (for example systematic review of many different studies evaluating the effectiveness of an intervention applied in similar but not exactly the same contexts). In the meta-evaluation of a mega-event, however, it is data from different subcomponents of the same event that needs to be synthesized, to provide a fuller understanding of the efficacy of the event and and/or its legacy.

Most of the papers in this review of the meta-evaluation literature provided only limited discussion of specific technical issues. The papers reporting specific meta-evaluation studies also provide little details of their methods. The result is that the literature is rich on conceptual issues, though no paper is comprehensive, but thin on technical issues.

Nonetheless, many of the papers are concerned with basic **standards** and stages of evaluation and sources of error. For example, programme evaluation standards have been produced that list criteria for evaluations and meta-evaluations for utility, feasibility, propriety, accuracy and accountability. These criteria and accompanying guidance are also of great value to the meta-evaluations of mega events given the wide range of academic and grey literature that such events tend to generate, and which needs to be sifted and appraised. The quality assessment of component studies is also an integral element of the synthesis; it can help weight and strengthen the claims made by the meta-evaluation of a mega-event. In the case of the Meta-Evaluation of the 2012 Olympic and Paralympic Games, which incorporates both formative and summative stages, multiple purposes for appraisal are present, including:

- The quality appraisal of methodological plans and activities to provide feedback for planned or ongoing constituent studies (for example to help align research objectives, and to ensure minimum standards of quality);
- Potentially a meta-appraisal of the state of research activity across whole meta-evaluation themes or sub-themes (i.e. to inform judgements on the extent to which devising and later answering specific research questions is viable); and
- An assessment of the relevance and trustworthiness of results from interim and final evaluations, and the related weighting of evidence to determine its 'fit for purposeness' for incorporation into the review.

The methods used to (meta) evaluate a mega-event can also follow some of the methods of systematic review. There is a very rich detailed literature on systematic reviews which form part of some definitions of meta-evaluation (see Gough et al 2012). The literature review in Section 3 would not have identified all of these studies as the search strategy was primarily aimed at studies describing themselves as meta-evaluations, rather than the very much larger literature on systematic reviews (i.e. an artefact of the search strategy). However section 3.4 does set out the broad types of review, which could be considered relevant to synthesis meta-evaluation. In the evaluation of mega-events, the event is so large and may have so many different aspects of interest, that it is likely that there will be a range of questions to be asked and thus many sub-reviews with different systematic review and synthesis methods that can be combined to address one or more overarching questions (St Pierre 1982).

Other papers identify sources of poor meta-evaluation practice, from factors such as inappropriate problem formulation, lack of independence of the meta-evaluators from the primary evaluations under study, poor quality of meta-evaluations and little monitoring of quality standards, which provide further useful hints for the meta-evaluation of mega-events.

Section 5 builds on these findings from the literature review, and methods of systematic review more widely, to provide a set of guidelines for structuring the methodology of the Meta-

Evaluation of the 2012 Olympic and Paralympic Games, as well as other impact meta-evaluations.

4 Analysis of expert interviews

4.1 Introduction

The initial review of the literature on meta-evaluation concluded that there are very different understandings of what constitutes meta-evaluation and a wide range of different 'meta-evaluation' methods in use. To explore these issues in more detail we undertook a series of semi-structured interviews with experts in the field. This report analyses the views of the experts who we consulted:

- The next section provides brief details of the backgrounds of the interviewees;
- Section 4.3 reports their views on what meta-evaluation is;
- Section 4.4 describes their assessment of the current state of the art of meta-evaluation and the main challenges which it faces;
- Section 4.5 presents the experts' views of how one might evaluate the legacy of the 2012 Olympic and Paralympic Games; and
- Section 4.6 draws together the key points to emerge from the interviews.

4.2 Interviewees

The interviewees are acknowledged experts in the fields of evaluation and/or sports policy. The initial sample of potential interviews was identified from the literature review (discussed in section 3.0) and the authors' own knowledge of the field. Thereafter a 'snowball' method was used which involved asking early interviewees to suggest others who they believed would have useful insights into meta-evaluation approaches.

A total of 18 experts were approached. Five declined to participate (some claimed not to know enough about meta-evaluation; one was unwilling to disclose details of the methods which they used). A total of 13 experts drawn from academia and consultancy firms and from across the US, UK and the rest of Europe were interviewed (see Annex 1). All 13 had direct experience of meta-evaluation research (broadly defined) or related activities such as meta-analysis.

Interviews were conducted using a topic guide which was adopted by all interviewers (see Annex 2). Results were recorded in contemporaneous notes taken by interviewees and analysed using a standard matrix.

4.3 Definitions

(i) Meta-evaluation in theory

The literature review undertaken as part of the methods development study identified three main schools of thought about what constitutes meta-evaluation.

Some researchers and commentators see meta-evaluation as being concerned primarily with standard setting. Seen in this light meta-evaluation is a process of establishing criteria for the evaluation of evaluations. The purpose of a meta-evaluation is to examine other studies against an established set of standards and goals in order to determine whether they were conducted in a rigorous and robust fashion.

A second school of thought sees meta-evaluation as a form of meta-theory. According to this view, meta-evaluation is concerned with the role of evaluation. It focuses on questions such as what can (and cannot) be evaluated; how (un)certain is the evidence; the extent that findings are transferable; and how we should understand causation in policy analysis.

A third strand of the literature describes meta-evaluation as an activity which brings together data and/or findings from a range of studies of initiatives or programmes to investigate overarching themes or draw out broader lessons for policy. This brand of meta-evaluation is concerned with retrospective holistic assessment of interventions. The aim is to identify repeat patterns and collective lessons across groups of similar policies or initiatives that have been implemented in different settings and/or at different times. This variant of meta-evaluation has much in common with systematic review and meta-analysis in that all three types of enquiry seek to bring together evidence to assess the efficacy of groups of programmes. But unlike systematic review or meta-analysis, which are identified by the particular methodologies that they employ, meta-evaluation is not linked to any particular kind of methodology or data. It is defined much more broadly and covers a wide range of different approaches and different type of study.

(ii) Meta-evaluation in practice

The interviewees confirmed several of the main findings of the literature review. There was wide agreement that the term meta-evaluation is a confusing one because it is used in very different ways by different scholars and practitioners. It was striking that some of the experts did not recognise the term meta-evaluation. Two of those who we approached declined to be interviewed for this reason and one experienced evaluator who had undertaken several 'meta-evaluations' told us that:

'There are such huge variations in meta-evaluation that it is difficult to say anything about what it is.' (Interviewee A)

There was near universal agreement among those who were familiar with the term meta-evaluation that there was very little meta-evaluation going on and that there was a need for much more of it. One noted:

'There are only one or two teams doing it in France. But it is needed.' (Interviewee L)

However, opinions about what meta-evaluation actually is were split roughly equally. Four interviewees were firmly of the view that it is about setting standards or judging the quality of evaluations. Six saw it as a process of synthesising the results of other studies as meta-evaluation in order to make judgements about the effectiveness of policies. Two believed that it has a dual function, combining both standard setting and synthesis.

(iii) Meta-evaluation as standard setting

One interviewee explicitly rejected the notion that meta-evaluation should be a process of standard setting, criticising:

'studies that restrict themselves to a small number of highly quantitative data from RCTs and exclude other evidence because of quality assurance concerns, leaving themselves just a small number of residual studies to draw on.' (Interviewee J)

The US experts interviewed were both firmly of the view that meta-evaluation was concerned with standard setting. One defined meta-evaluation as 'a process or summative evaluation of the technical quality of evaluations'. The other was engaged in processes of capacity building and quality assurance of the work carried out by evaluators. Their agency has established standards relating to the way in which data are presented and causality is demonstrated. Evaluators submit their proposed methodologies for examination against these standards which are seen as providing 'a quality benchmark'. But in addition to assessing evaluations, the agency also provides technical advice to evaluators and financial support to those policy makers to assist them in designing evaluations of interventions.

Some of the British and French experts agreed that quality assurance was part of meta-evaluation but emphasised the importance of capacity building as opposed to standard setting. One reported on their experience of having acted as scientific adviser to a UK Government department on a programme of 12 evaluations of related policy initiatives over a period of several years. This work involved assuring the Department that the evaluations were being conducted in a rigorous way (standard setting) and identifying the overall findings that emerged from the programme of work (synthesis). They had become closely involved in advising the 12 evaluations on methods and acting as what they described as a 'go between' between the evaluation teams and the Department funding the work. In their view this kind of 'hands on' approach was crucial to the success of both aspects of their meta-evaluation. Working closely with other evaluators to enhance the quality of the evaluations was, they argued, the best way to gain access to the data which were needed from these projects by the meta-evaluation in order to enable it to provide an overall assessment of policy. In their view:

'Meta-evaluation needs to talk with the other evaluations. It's not so much about methods as about management.' (Interviewee E)

Another expert spoke of the role of meta-evaluation in shaping expectations of what evaluations can be expected to deliver. They believed that the terms of reference issued by commissioning bodies are often too ambitious. By looking back at what studies have actually been able to achieve meta-evaluation could help to produce more coherent and consistent terms of reference for future studies.

(iv) Meta-evaluation as impact assessment

Several interviewees emphasised that meta-evaluation should make a positive difference. For three experts its primary purpose was to improve policies by analysing the impact and effectiveness of groups of evaluations. Three others saw meta-evaluation as being concerned primarily with improving evaluations. For them meta-evaluators should not just set standards but must also help to improve capacity by providing support and advice to evaluators in order to conduct better studies. Two of the European experts from outside of the UK saw meta-evaluation as the study of impact and use made of evaluations. One described it as the:

'evaluation of the effectiveness and impact of evaluations.' (Interviewee G)

The other as:

'Impact assessment of evaluations reports on the policy processes and decisions.' (Interviewee L)

They advised that this was the commonly understood definition of meta-evaluation in European evaluation circles. Methods for this kind of study were, they said, well understood and were presented to European evaluation standards.

(v) Meta-evaluation as synthesis

Half of the interviewees described meta-evaluation as a process of synthesising evidence from other studies. One encapsulated this view:

'an overarching evaluation which draws together a range of studies to reach overall conclusions.' (Interviewee J)

Another defined meta-evaluation as:

'A research method for evaluation of large programmes based on existing evaluations.' (Interviewee K)

Several interviewees noted that meta-evaluation takes a broader and longer term perspective than other forms of evaluation. They described meta-evaluations as focusing on 'overarching' themes or impacts and taking a longitudinal approach. They argued that by taking a more 'holistic approach' meta-evaluation was able to:

'understand the higher level mechanisms that are not visible from the secondary sources' (Interviewee K)

and the complex interactions between policies:

'the synergies between programmes that make the total effect greater than the sum of the individual parts.' (Interviewee K)

Several interviewees had conducted studies that sought to synthesise evidence from evaluations of groups of related policy initiatives or programmes. Some had led national evaluations which drew data from studies of local projects or partnerships to provide overall assessments of their impacts on 'high level outcomes' such as worklessness, quality of life, health and educational attainment. Others had been responsible for studies which brought together data from a range of national evaluations to reach an overall assessment of international aid programmes. Two experts from the rest of Europe recognised this kind of activity but described it as synthesis rather than meta-evaluation. For them synthesis was:

'evaluation of a programme, based on exclusively other evaluations' (Interviewee G)

'an evaluation primarily based on other evaluations' (Interview L)

However some of the other interviewees disagreed. For one, meta-evaluation:

'is the aggregation of broadly similar outcomes by bringing together different studies and different types of evidence.' (Interviewee K)

whilst synthesis involves:

'the aggregation of broadly similar types of evidence about broadly similar kinds of outcomes.' (Interviewee J)

Another suggested that meta-evaluation draws exclusively on other evaluations whilst synthesis uses databases and other secondary sources alongside the findings of other evaluations.

Other interviewees noted the similarities in terms of objectives but differences in terms of methods between meta-evaluation and systematic review and meta-analysis. All three activities were concerned with what one called 'a review of study results'. However meta-evaluation is 'a broader concept than systematic review which has formal rigour and gravitates towards quantitative studies', whilst meta-analysis is more narrowly defined still. It is 'a statistical toolkit that enables you to assimilate very specific sets of data in very specific ways using regression analysis.'

4.4 The state of the art

Having established how they defined meta-evaluation, the experts were then asked for their views on the current state of the art – its strengths, weaknesses and the main challenges which meta-evaluators must confront.

(i) Standard setting

Those who regarded meta-evaluation as standard setting reported that it was a well-established and well regarded activity. There were clear sets of criteria and established methodologies that are widely used (as detailed in our review of the literature), and assessments were generally rigorous and useful. They reported that in the US, where this type of meta-evaluation is most prevalent, the emphasis had traditionally been on ensuring the quality of evaluation designs. However there has been a growing realisation that good design is not a guarantee of good evaluation. Implementation matters as well. The US Government has therefore paid increasing attention to the ways in which evaluations are conducted.

Approaches to monitoring have included the appointment of expert working groups and recruitment by government agencies of staff with expertise in evaluation methods. Interviewees reported that technical working groups are good in theory and often work well, although some lack the necessary expertise or are captured by a few influential members.

(ii) Meta-evaluation as synthesis

Those who saw meta-evaluation as synthesis of the results of other studies were enthusiastic about its potential. Policy agendas are complex and ambitious. Policy makers look to interventions to produce massive changes (such as health service modernisation), deliver on heroic targets (such as reducing levels of worklessness) or serve grand philosophies (increasing volunteering in the big society). Initiatives inevitably interact with each other. Some are designed to be mutually reinforcing; others may unintentionally cut across one another.

In recent years there has therefore been growing interest in whether policy-making is 'joined up'. Rather than studying projects, programmes or policies in isolation, it makes sense therefore to adopt a holistic approach which examines their collective impact. And interviewees argued that longitudinal studies that seek to identify 'higher level' outcomes and the interactions between policies should be more efficient than evaluations which focus on narrowly defined policy agendas and more immediate impacts.

Interviewees reported a number of advantages over other forms of evaluation research:

- **Longer term trends** – Because many meta-evaluations are longitudinal studies, they enable researchers to recognise trends which go beyond specific interventions. Speaking of a large, 10-year meta-evaluation that he had led, an interviewee reported that:

> 'The huge benefit was ability to study change over time in way most evaluations can't get at' (Interviewee F)

- **Repeat patterns** – Meta-evaluation can help to reiterate lessons from the past which policy makers may easily have forgotten. As one interviewee put it, meta-evaluation:

> 'Can keep lessons of evaluations alive; many times the learned lessons from an evaluation of 3-4 years ago are already forgotten.' (Interviewee G)

- **Influence** – Meta-evaluation may also gain more attention than studies of individual interventions. It is:

> 'a great tool for programme managers to steer the programme' (Interviewee K)

And it is:

> 'more likely to reach a target audience high up in the hierarchy of the commissioning organisation, as it summarizes other evaluations' (Interviewee G)

- **Cost** – Although meta-evaluation studies tend to have large budgets they may be more efficient than other forms of evaluation because they use existing evidence. They help to give:

> 'added weight to evaluations that are included' (Interviewee G)

And this enhances:

> 'the value of existing evaluations' (Interviewee K).

(iii) Theory and methods

In spite of their endorsement of and evident enthusiasm for meta-evaluations which seek to synthesise evidence and data from other sources, interviewees noted that in practice there have been very few studies of this kind (an observation which is borne out by the review of the literature). There is no established theory of meta-evaluation. And in contrast to the practice of meta-evaluation as standard setting, the literature on meta-evaluation as synthesis is underdeveloped. One interviewee told us:

'As far as I know there is no written material. There are no benchmarks or rules, no knowledge platform .....It should be possible to design general rules that are harmonious with all (studies).' (Interviewee G)

Another believed that part of the problem was the lack of training in methods:

'We just don't have enough evaluators from evaluation schools.' (Interviewee L)

Meta-evaluation methods borrow from other branches of evaluation research and the social sciences in general. But typically each meta-evaluation is designed from scratch:

'The wheel is reinvented over and over. There is not enough transfer of knowledge between meta-evaluation experiences.' (Interviewee K)

These problems are compounded by the complexity of the issues which meta-evaluations are often seeking to address. Whilst in theory one of its major attractions is the focus on groups of policies or interventions, in practice it can be very difficult to model and measure interactions between them.

Two interviewees argued however that the problem was not a lack of good theoretical frameworks or methodological templates, but a lack of confidence in using what was already available. One argued that meta-evaluation could make use of theory-based evaluation, contribution analysis[8] and realist synthesis (covered in the review of the literature). Another commented that:

'There are some good meta-evaluation designs but they are rarely implemented in practice because sponsors and consultants want simpler frameworks ..... You watch your advice being ignored by funders - partly through fear that this will throw up unwelcome findings. So they give it to a safe pair of hands to do the work, consultants who go back into conventional methods like surveys and case studies because that's what the Department wanted.' (Interviewee J)

(iv) The politics of meta-evaluation

Three interviewees spoke of the politics of meta-evaluation. They suggested that because meta-evaluation addresses high profile policy objectives and 'flagship' programmes, the stakes are often higher than for more narrowly defined evaluations. This makes meta-evaluation more visible which can enhance the prospects of utilisation. However, they reported that their own studies had run into problems with funders when the findings suggested that interventions had not had the significant effects that policy makers had hoped for.

(v) (Accessing and aggregating) secondary data

According to some of the experts, its reliance on evidence and/or data collected by other evaluations is one of the defining features of meta-evaluation, marking it out from other forms of synthesis. And many of the interviewees saw its ability to aggregate different kinds of evidence and data as one of its main attractions. But they also acknowledged that in practice it

[8] See for example: http://www.scotland.gov.uk/Resource/Doc/175356/0116687.pdf

could be difficult to access and then use secondary data. Synthesising data is, one said, 'a primitive art'.

Some interviewees with first-hand experience of trying to synthesise evidence from other evaluations reported that they had found it difficult to persuade other evaluations and stakeholders to share data. Others told us that when they were given access to the evidence collected by other studies, it was not very useful for their meta-evaluations because it tended to be often focused on narrowly defined policies and outcomes. They also reported problems assimilating data that had been collected for different purposes, by different teams, at different times, using different samples and methods. In light of this experience one interviewee concluded that:

'The greatest problem for any meta-evaluation is the heterogeneity of the data it uses.' (Interviewee A)

Another agreed:

'The biggest challenge is the problem of incommensurability. You are usually trying to build in retrospectively a coherence that wasn't there prospectively'. (Interviewee K)

A third said that it was vital to:

'make sure all individual evaluations use the same yardstick to measure outputs on.' (Interviewee K)

An interviewee who specialises in meta-analysis explained that it can only use very specific types of evidence: quantitative data (preferably expressed as a 'real number') and multiple, similar, replicable datasets (the more observations the greater the reliability of the analysis). Conversely, experts in meta-evaluation agreed that given the shortage of available data, they can generally not afford to be this selective. One was especially critical of studies that restrict themselves to:

'highly quantitative data from RCTs leaving lots of evidence out because of quality assurance concerns and leaving a small number of residual studies.' (Interviewee J)

But others doubted the feasibility of synthesizing the results of evaluations that were not experiments.

Interviewees suggested three practical steps which could help alleviate problems relating to secondary data. First, they suggested that the sequencing of meta-evaluation and other studies is important. Many meta-evaluations are commissioned after the studies upon which they were supposed to draw. As a result they have very little, if any, influence over what data are collected. And those undertaking the other evaluations may see the involvement of meta-evaluators as an unwelcome complication and added burden on them. Commissioning the meta-evaluation first would mean that the meta-evaluators could be involved in the design of other studies in order to ensure that they provided data which could be synthesised.

The interviewees' second recommendation was that a requirement to work with a meta-evaluation should be written into protocols and contracts agreed among the funders, meta-evaluators and the other evaluation studies.

Third, they said that it is important for meta-evaluators to build a rapport with other evaluations on which they could draw by assisting them in their tasks. Two of the experts reported that in the course of meta-evaluations which they had conducted they had spent a lot of time helping the other evaluators to develop their evaluation methods, identify common themes, negotiating data sharing protocols etc. As one put it:

'you need to try to add value for the individual evaluations as well as sucking out value for the meta-evaluation .... 'You've got to talk to people throughout the process, not just when they are designing studies or reporting their findings ..... You need to be a fly on the wall not a fly in the ointment'. (Interviewee E)

(vi) Attribution

Some of those who had conducted meta-evaluations reported that their studies had failed to detect significant changes in the higher level outcomes which they had focused on. Sometimes this was because it was difficult to establish a credible counterfactual. Studies had lacked baselines against which change could be measured or had had to use a series of 'ragged' baselines (i.e. different baselines for different policies). This made it difficult to know what point in time to track change from. But even where there were reasonably good baselines, policies had often apparently failed to have much of an impact. This is not too surprising given that the meta-evaluations were said to be often focused on 'wicked issues', which had proved largely immune to previous interventions. However, this was not what policy makers wanted to hear and could make for a difficult relationship with the funders (see section 4.3 above).

Where there were changes as a result of interventions it is often difficult for meta-evaluations to establish attribution because of the wide range of factors which could have influenced outcomes. Establishing cause and effect is a problem for all evaluative activity. However, interviewees said that the challenge was particularly acute in the case of meta-evaluation because it tends to focus on high level, longer term objectives which are likely to be affected by a wide range of policies and other influences.

One expert summed it up as follows:

'Although the work process might be similar to other evaluations, the work field is much more complex. It is difficult to prove or even understand cause-effect processes. And the evidence is very anecdotal. It is more based on words, discourses..... which makes it more biased as the proportion of facts is low.' (Interviewee L)

Those who saw meta-evaluation as being concerned with assessing the impact of evaluations reported similar difficulties. They observed that it was very difficult to work out how a policy had originated and to establish a link with particular studies.

4.5 Implications for the evaluation of the 2012 Games

Turning to the Meta-Evaluation of the 2012 Olympic and Paralympic Games, the experts were asked how they would approach this task and in particular what methods they would recommend for integrating evidence from other studies and datasets.

All of them agreed on the need to first determine what questions the study needs to focus on. They believed that the starting point should be discussion and agreement about:

- What is meant by the concept of legacy in the context of the Games (including the important question of legacy for whom);
- What the mechanisms for achieving this legacy are; and
- What data will be available to meta-evaluators?

Only then could the question of methods be addressed.

Interviewees emphasised the value of focusing on 'high level' themes. Several recommended developing an 'overarching framework' which modelled the intended outcomes (improvements in the economy, social capital, the environment, etc) and the more specific mechanisms associated with the Games that might reasonably be expected to contribute to these legacies. The framework should, they said, also identify potential interactions between the different types of legacy and between different mechanisms.

There was a measure of agreement about what the 'big' themes should be. Almost all of the interviewees recognised the importance of the economic legacy of the games and its impact on the environment and sports participation. Some argued that it was also important to consider the 'political' or 'governance' legacy – for example the impact of the Games on relations

between the 'Olympics' boroughs in which they are situated. There were differences of view about the notion of social and cultural impacts. Some believed that they are an important component and there are examples in the literature of evaluations which include these impacts, but others argued that these were too ill defined to be included. One interviewee said that he would steer clear of cultural impacts because they were:

'very soggy and not well researched in previous studies.' (Interviewee A)

The same interviewee argued that the meta-evaluation should also use some overall measures of legacy such as 'well being' or 'quality of life'. He claimed that progress had been made in recent years in measuring citizen and staff satisfaction in public service organisations and noted the UK Government's interest in measuring 'happiness'. It might, he suggested, be possible to revive the (recently abolished) Place Survey in the Olympic Host Boroughs in order to track changes in local peoples' satisfaction with public services and their perceptions of these areas as places to live.

Several experts recommended a theory led approach as a means of constructing such as model. One described this process as:

'specifying the pathways to the impacts.' (Interviewee A)

Another advocated what they called a 'content based approach based' which:

'iteratively builds a model of the scope, content, and possibilities of the various types of legacy you want.' (Interviewee ?)

The resulting framework could, they suggested, be used to:

'help to look for similarities in the mechanisms and then have conversations with the other evaluations about the data which they can offer.' (Interviewee J)

They anticipated that some aspects of the Games would have important impacts on several different kinds of legacy and that the meta-evaluation might therefore want to prioritise and focus on these.

In a similar vein, another expert recommended:

'The use of logical frameworks and questioning programme managers about where they think the project fits within the whole of the programme, to reveal interdependencies.' (Interviewee K)

Another advocated what they called:

'Screening and scoping - done in iterative steps and with an exploratory phase if required - to improve the interdependency matrix and assumptions about cause-effect chains.' (Interviewee L)

The experts also emphasised that in their experience it was important for a meta-evaluation to work closely with other evaluations from which useful data might be obtained. One said that once the key questions for the meta-evaluation had been defined, it would be important to:

'have conversations with other studies and map their contributions ..... to see the overlaps and the gaps in the data that will be available to the meta-evaluation.' (Interviewee J)

Another suggested an alternative (or perhaps complementary) approach to identifying impacts based on drawing:

'a sort of Venn diagram which looks at the four (or however many) themes you have and the data which will be available from other sources.' (Interviewee E)

The experts also recommended testing the robustness of the studies which the meta-evaluation might draw upon. One advocated a method based on sampling of conclusions and testing the strength of the evidence base which underpinned them. He suggested that studies should then

be ranked in terms of their reliability and the results of those rated as good should be weighted more heavily than those about which there were concerns.

Several interviewees argued that it will be important to evaluate variations in legacy impacts – over time and over space. One interviewee distinguished between 'immediate impacts' (effects that were evident before, during or soon after the Games but were not expected to last in the longer term); 'sustainable impacts' (effects that persisted for some time after the Games); and 'generative impacts' (effects that in turn created further benefits (or dis-benefits) – for example, the multiplier effects associated with regeneration facilitated by the Games. They recommended that the meta-evaluation team:

'Engage with stakeholders who will 'enact legacies'. They might for example convene a group of 'legacy inheritors' because sustainability is important.' (Interviewee J)

Several of the experienced evaluators to whom we spoke to however cautioned that the stated objectives of the meta-evaluation seemed over ambitious. They had particular concerns about the concept of a counterfactual because of the range of other factors that will affect regeneration, employment, health and sports participation and so forth. One argued that it would be:

'Impossible to know in a recession what the counterfactual would have been because you can't just compare to previous years.' (Interviewee G)

4.6 Conclusions from the expert interviews

The interviews with some of the leading experts in the field of evaluation provide some important pointers for the Meta-Evaluation of the 2012 Olympic and Paralympic Games.

They confirm some of the main findings of the literature review. They show that there is considerable confusion surrounding the term meta-evaluation. Some experts are unaware of it. Others are familiar with it and regard it as important but have quite different views of what meta-evaluation actually entails. Opinion is divided into two main camps: those who see it as a process of judging the quality of evaluations; and those who regard it was a way of judging the effectiveness of policies or programmes.

The implication is that it is important to **be clear about the purpose of the meta-evaluation** of the 2012 Games. This places the Meta-Evaluation of the 2012 Olympic and Paralympic Games firmly in the synthesis camp. It will draw on evidence from a range of sources including other evaluations and therefore does need to demonstrate that secondary data are reliable. However, the primary task is to provide an overall assessment of the effectiveness of the Games in delivering a legacy, rather than on the rigour of other evaluations.

Second, the experts believe that in order to provide this overall assessment it is necessary to **define the nature of the legacy which the Games are intended to achieve.** In practice there are likely to be a number of different types of legacy. The experts suggested that at the very least the Meta-Evaluation should consider economic, social, environmental and sporting legacies. They also pointed to a number of other potentially important impacts, including the political and governance legacy (for example for East London).

Third, the interviews revealed that as well being clear about the type (or types) of legacy **it is** important to **be clear about the distribution of the legacy**. This means that the meta-evaluation will need to try to assess which areas and which sections of society benefit (or experience dis-benefits) from the 2012 Games.

Fourth, it will be important to know not just whether but also **how the legacy is achieved**. Several of the experts recommended developing a theory based approach which models the ways in which the Games might lead to legacies and then tests whether these have occurred in practice.

Fifth, several experts were clear that one of the main benefits of meta-evaluation is that it encourages a 'holistic' assessment of groups of policies or programmes. This implies that the meta-evaluation should **focus on 'high level outcomes' and pay attention to interactions** between different aspects of the Games and potential synergies between different types of legacy.

Sixth, most interviewees identified problems concerning data availability. Those who specialise in specific techniques (for example meta-analysis) that require particular types of data advised that their methods could not be easily applied to the meta-evaluation of the Games because the data are unlikely to be available. For this reason the meta-evaluation of the 2012 Games will need to take a pragmatic approach which **draws upon a range of very different kinds of evidence**, but also looks to **work with and if possible influence component evaluations**, as well as appraising their relevance and quality**.**

Finally, most of the experts believe that meta-evaluation is necessary and worthwhile but they caution that it presents formidable methodological challenges. As with many other complex interventions, it will be difficult to identify clear baselines or counterfactuals for the 2012 Games. Establishing cause and effect mechanisms will not therefore be straightforward, and time lags may well mean that the full extent of any legacy is not measureable within the time frame of the study. For these reasons it is important to have **realistic expectations of what can be achieved** and to focus the meta-evaluation effort on those issues which are most important, and for which evidence is available.

5 Guidelines for meta-evaluation

5.1 A framework for conducting impact meta-evaluation

It has been identified that the meta-evaluation of mega-events most closely resembles the synthesis of evaluations form of meta-evaluation. An early step for the Meta-Evaluation of the 2012 Olympic and Paralympic Games should therefore be to determine how relevant evaluations and their results are to be identified and integrated, in response to the overarching research objectives. This requirement, and particularly once broken down into its constituent parts, would appear to have the characteristics of a multi-component, mixed-methods systematic review. The wide variety of impacts identified however means that the specific type of data sought, the appraisal criteria deployed and the methods for synthesising the data will differ widely across the Meta-Evaluation.

Nonetheless at its heart the Meta-Evaluation of the Olympic and Paralympic Games is driven by a set of logic models hypothesizing how the Games might impact on four broad types of outcome. Such a theory driven evaluation could benefit from some of the aims, methods and organising principles of realist synthesis, discussed in the previous section. This would involve seeking to test the efficacy of groups of interventions and phenomena relating to the Games' legacy against a pre-constructed (but nonetheless malleable) set of assumptions or programme theories, concepts and measures, based upon data collected in a relatively systematic way.

Pawson et al (2004) mapped out the process for undertaking realist synthesis (as one form of theory driven synthesis). Combining this approach with the steps taken in systematic reviews provides a useful starting point for establishing a process for conducting impact meta-evaluation as shown in Figure 3 (the process may be iterative but is shown as a linear list here for clarity). Although this has been designed to help structure the methodology for the Meta-Evaluation of the 2012 Olympic and Paralympic Games, it also has more universal applicability.

Figure 5-1: Stages of an impact meta-evaluation: a linear list of an iterative process (informed by Pawson et al 2004)

1. DEFINE THE SCOPE OF THE META-EVALUATION
   - 1.1 Identify the purpose of the meta-evaluation
   - 1.2 Clarify aims of evaluation in relation to theory testing
   - 1.3 Clarify theories and assumptions
   - 1.4 Design an evaluative framework to be populated with evidence

2. IDENTIFY STUDIES
   - 2.1 Clarify information required
   - 2.2 Develop strategy to identify this information
   - 2.3 Develop methods to identify this information
   - 2.4 Screen to check that information identified fits information required
   - 2.5 Compare available information against what is required
   - 2.6 Consider seeking further information

3. CODING FROM STUDIES

Develop strategy and methods to collect information from studies in order to:

   - 3.1 Manage information through the review process (e.g. using data extraction templates)
   - 3.2 Ensure that it meets evidence needs of review/evaluative framework
   - 3.3 Map the information

3.4 Enable quality and relevance appraisal
3.5 Provide the information to enter into the synthesis

4. QUALITY AND RELEVANCE APPRAISAL

Develop strategy and methods to assess the:

4.1 Rigour by which the information has been produced
4.2 Relevance of the focus of the information (such as intervention, context, outcomes) for answering the review questions or sub-questions
4.3 Fitness for purpose of the method by which the information was produced for answering the review questions or sub-questions
4.4 Overall weight of evidence that the information provides in answering the review questions or sub-questions

5. SYNTHESIS

Develop strategy and methods to:

5.1 Clarify the evidence available for answering the review questions and sub-questions
5.2 Examine patterns in the data and the evidence they provide in addressing review questions and sub-questions
5.3 Combine evidence from sub-questions to address main questions and cross cutting themes
5.4 Test the robustness of the syntheses

6. CONCLUSIONS AND DISSEMINATION

6.1 Engage with users of the meta-evaluation to interpret draft findings
6.2 Interpret and test findings
6.3 Assess strengths of the review
6.4 Assess limitations of the review
6.5 Conclude what answers can be given to questions and sub-questions from evidence identified
6.6 Refine theories in light of evidence
6.7 Disseminate findings

In this way, the synthesis includes empirical outcome data. It also includes generating, exploring and refining theories of process, including what works, for whom, in what contexts and why (and the interactions between interventions), based on more iterative methods and qualitative forms of synthesis to configure such findings as systematically as possible from the available evaluation evidence (and to help interpret outcome data). The latter could also include elements of the evaluation where 'cause-and-effect' analysis is less appropriate (e.g. for complex adaptive systems) or where levels of uncertainty in some areas of analysis makes cause and effect analysis of little use.

5.2 Stages of an impact meta-evaluation

(i) Stage 1: DEFINE SCOPE OF THE META-EVALUATION

Step 1.1: Identify the purpose of the meta-evaluation

- Type/nature of the intervention(s)
- Overall policy or other aims that may be achieved
- Specific impacts
- Context for these policy aims and specific impacts to be achieved

Step 1.2: Clarify review aims of evaluation in relation to theory

- Integrity: does the intervention work as predicted?
- Comparison: what is the relative effect for different groups and settings
- Adjudication: which theories best fit the evidence?
- Reality testing: how does the policy intent translate into practice?

Step 1.3: Clarify theories and assumptions

- Search, list, group, and categorise relevant theories (configurative synthesis)

Step 1.4: Design an evaluative framework to be populated with evidence

- Specify review questions and sub-questions
- Specify review methods for questions and sub-questions

The literature on mega-events highlights the importance of the research questions being addressed through the meta-evaluation, the direct and indirect indicators to be used to address these questions, and the time span over which the questions are to be considered. In all cases, the types of primary research and data considered for inclusion in the meta-evaluation, the methods to quality and relevance appraise data from those studies, and the methods used to synthesise the quality and relevance appraised data will depend upon the nature of each question being asked. The questions can be multiple and complex and at many different levels of analysis and of more or less concern to different stakeholders.

The questions asked will firstly depend upon the broader user perspectives and interests of those asking the questions. The first step in a meta-evaluation (and in all research) is to ensure clarity around why the evaluation is being undertaken, for whom and for what purpose. In other words, who are the users of the meta-evaluation? Different individuals and groups will have different interests and thus different questions and these questions will contain theoretical and ideological assumptions of various types. In this way, user perspectives drive the specification of meta-evaluation questions (and this will in turn mean the analysis and reporting of some elements of the evaluation from different stakeholder perspectives).

The questions being asked by meta-evaluations of this type also concern the evaluation of interventions. However these meta-evaluation questions are not necessarily the same as the questions addressed by the individual studies (and may not treat the data of the studies in the same way as the individual studies do). Rather, meta-evaluation research questions will be tested within the specific circumstances and context of particular, often over-arching policy aims and objectives. Even if seemingly framed as generic questions, the questions will be asked in relation to specific policy goals in specific social and material contexts. The meta-evaluation then interrogates each included study to determine the extent that it helps to address each specific meta-evaluation question. Moreover, these larger macro questions must be addressed by asking sub-questions relating to more specific and often narrower examples of the generic intervention and/or more specific and often narrower outcome measures. The meta-evaluation thus becomes a synthesis of sub-questions and sub-studies to address the overall meta-evaluation question.

As all questions and stakeholder interests cannot be addressed there has to be a process for the identification and prioritization of specific questions. The implicit theoretical and ideological assumptions (sometimes called the conceptual framework) need to be explicit to assist the process of prioritization (and the specification of review methods). This can include the modelling of the processes (or mechanisms) by which positive or negative outcomes are thought to occur (sometimes called logic models), and the relationships between overall questions and models and the various sub-questions and their models.

The research question then becomes one of assessing the impact of the mega-event and is essentially the testing of a hypothesis of a 'theory of change' (or multiple sub-theories of change). The starting idea is that the event will have some positive (and maybe some negative) effects. The preliminary idea, ambition, expectation, hypotheses or 'programme theory' is that if

certain resources (material, social, cultural) are provided to deliver the mega-event then those resources will engender individual behaviour change and community action to a sufficient extent that benefits will follow and a lasting legacy will remain. Like all hypotheses, these speculations turn out to be true or false to varying degrees.

These eventualities provide the underlying logic for theory-driven evaluation. Research begins by eliciting the key theories assumed in the construction of programmes and then goes on to test their accuracy and scope – the programmes is supposed to work out like this but what happens in practice? Empirical inquiry is conducted with the task of discovering where the prior expectations have proved justified or not and can involve analysing 'process', 'outputs' and 'outcomes', as specified for example in the logic model. This in turn can involve a multi-method approach employing qualitative, quantitative, documentary, comparative and retrospective inquiry, but which will differ according to each specific research question.

Any intervention is nonetheless likely to have differential effects if provided in different ways to different groups in different situations; the theories can also be tested to assess the extent that they can predict such variation. There may also be a variety of theories that attempt to explain the effects of an intervention and the meta-evaluation can aim to assess the relative strength of the theories in predicting effects (if this is one of the review aims). Finally, the logic models and theories need to be sufficiently flexible to take account of more innovative or 'generative' interventions (for example 'learning by doing strategies'), and the emergent nature of any outcomes generated.

The research strategy for this type of meta-evaluation can therefore be no better than the concept maps which commence it. The 'theory elicitation' stage is crucial and formal review methods can also be used to identify and map these theories and concepts. The various theory and concept maps then need to be refined, through examining closely:

i) **Model verisimilitude and logic:** are the maps close enough to the working hypotheses of key policy architects?
ii) **Operational potential:** How feasible is the measurement and gathering of data on the processes and staging posts that are identified?

The greater the theoretical understanding of the issues to be tested, then the greater the specification of the research focus, rather than the 'black box' approach that is simply studying whether a difference is or is not associated with different experiences.

As briefly discussed in the preceding chapter, in the case of the impact meta-evaluation of complex government programmes and phenomena (such as a mega event), these logic models may need to be broken down along thematic lines, to help elucidate the detail involved in each theory of change. However this also needs to recognise the interactions and crossovers between different themes of activity, in that for example one aspect of a mega event or legacy investment may contribute to multiple outcomes and thus themes. This needs to be taken into account (through a theory rather then programmatic-led approach), and mechanisms for sharing knowledge established, to ensure that these synergies are not missed. If not then questions may be inappropriately formulated and appropriate data may not be sought by the impact meta-evaluation at the operational stage of the research, with the result that the full range of possible benefits (and potential disbenefits) may not be captured within each theme. Through developing the logic models and their accompanying theories of change, additional and important cross-cutting issues for the meta-evaluation may also emerge (for example issues of equality, effective process and sustainability/longevity) which can then be applied consistently across each theme (and sub themes) through the questions that are developed.

These maps then form the basis of an evaluative framework for considering meta-evaluation questions, informed by theories and concepts that provide the basis for both the specification and interrogation of evidence to answer these questions. As complex questions are likely to be too large to be studied in one go and need to be broken down into sub-questions (and maybe even sub-sub-questions), as well as cross-cutting questions, the meta-evaluation operates at multiple levels. These questions in turn inform the specific methods of meta-evaluative review

that need to be applied to identify, appraise and synthesize the relevant evidence. This process can be described as follows:

- Selection of overall meta-evaluation questions
  - Process of selecting stakeholders and involving in question selection
  - Criteria for selecting questions, cross-cutting questions and consideration of other questions not selected
  - The overall theoretical and ideological framework/complexity model of the questions being considered
  - The review methods used to identify, appraise and synthesize the relevant evidence

- Selection of sub-questions
  - Process of selecting stakeholders and involving in sub-question selection
  - Criteria for selecting sub-questions and consideration of other questions not selected, including how they answer the overall question and cross-cutting questions;
  - The theoretical and ideological framework/complexity model of the sub-questions and how they relate to each other and to the overall questions and framework including 'process', 'outputs' and 'outcomes'; and
  - The review methods used to identify, appraise and synthesize the relevant evidence.

Inappropriate problem formulation is a major risk. If the research question is not clear then it is unlikely that it will be operationalized in the research study in a way that the study will be able to answer it. Clarifying the purpose of the review, finding and articulating programme theories (and their interactions), and formulating meta-evaluation questions, sub-questions and cross-cutting questions should therefore constitute important elements of the scoping phase of the Meta-Evaluation of the Olympic and Paralympic Games.

(ii) Stages 2: IDENTIFY STUDIES

Step 2.1: Clarify information required

Step 2.2: Develop strategy to identify this information

Step 2.3: Develop methods to identify this information

Step 2.4: Screen to check that information identified fits information required

Step 2.5: Compare available information against what is required

Step 2.6: Consider seeking further information

The research questions and the associated evaluative framework drive the strategy for the search for and assessment of relevant evidence. As a meta-evaluation of a mega-event may ask very broad policy questions about, for example, the effects of the event on different outcomes, the process of clarifying sub-questions through 'surfacing' the logic models implicit in those questions will in turn help to clarify the type of data that will help assist in answering whether or not the interventions have had their hypothesized effects.

The evidence that is being sought can be described as 'inclusion criteria' and the extent that these can all be described a priori or develop iteratively will depend on the review strategy. The particular methods used to search for evidence that fits these criteria will similarly be framed by the particular methods of review being applied. The review method will determine whether the search for evidence aims to be exhaustive or not. Exhaustive strategies aim to avoid selection

bias by including all relevant data. Non exhaustive purposive strategies take a more iterative strategy of exploring investigative routes to test hypotheses. They aim for a more purposive and manageable analysis of discreet studies and/or sets of studies (and in some cases other secondary and primary data sets) to the extent that this facilitates the answering of different evaluation questions (and in the case of the meta-evaluation of a mega-event, in relation to specific components of impact and legacy)

The data that is available in practice is, of course, also limited by the studies available. In the case of the 2012 Games, this is likely to include primary evaluation studies set up specifically to evaluate Games components, other studies that happen to have been undertaken and are relevant, and ongoing and one-off surveys that may inform the synthesis. The studies may provide data on change subsequent to the Games, and/or data to provide evidence of additionality (to control for counterfactuals). There may also be many gaps. Relevant data sources (as well as gaps and potential contingency plans) can then be identified through stakeholder consultation and desk review (using methods for searching for studies developed for systematic reviews), and mapped against the research questions and indicators identified.

The data sources identified then need to be initially checked (screened) against the data required (inclusion criteria), and then consideration given to whether further data should be sought (and this may include the commissioning of further primary research). Searching for relevant data is seen as a step prior to quality and relevance appraisal of such data (considered later in this Section) but in practice these processes overlap as appraisal of fitness for purpose of identified data also relates to the need for searching for further data. A further complication is that a particular piece of data may have multiple roles and be applicable to varying extents in helping to answer more than one question and sub-question.

Importantly, this process also needs to be potentially undertaken at two levels (at least) for a meta-evaluation:

- For each sub-question
  - Specifying the information/data required to answer each sub-question
  - Scoping the information/data available or potentially available to be combined (synthesized) to answer the sub-questions (including checks for alternative explanations)
  - Identifying what further data are necessary

- For the overall or top level questions:
  - To what extent will the sub-questions provide answers to the overall questions?
  - Identifying what further data are necessary

(iii) Stage 3: CODING FROM STUDIES

Step 3.1: Manage information through the review process

Step 3.2: Ensure that meets evidence needs of review

Step 3.3: Map the information

Step 3.4: Enable quality and relevance appraisal

Step 3.5: Provide the information to enter into the synthesis

In a meta-evaluation or other form of review of primary studies, information will need to be recorded about each study. Some of this recording may use a priori categories and some may be open text coding. Both forms of coding have their advantages. Open text coding allows for a

richness of data but complexity in its analysis. Closed coding is much easier to analyse and arises from clear prior understanding of what is being sought from the coding.

In undertaking a synthesis of evidence there are at least five reasons for coding information from each study. The first is to describe the study in general ways to keep track of the study through the process of the review. A meta-evaluation is a complex process involving many questions and sub-questions and the identification of many pieces of data that may have multiple purposes in different parts of the analysis. There is thus an information management requirement for identification of data and their values in relation to different roles and stages in the meta-evaluation process.

A second reason is to provide information in order to assess whether the data meets the inclusion criteria for the meta-evaluation and thus be included in it. A third reason is to be able to describe (or 'map') the nature of the research field of research evidence meeting the inclusion criteria. It is also possible that not all the evidence included in the map will be synthesized, so that the synthesis is on a sub-set of studies from the map). Coding would then also be needed in order to select that sub-set of evidence. A fourth reason is to provide information to enable the quality and relevance appraisal of each piece of evidence to check whether it is fit for the purpose of the synthesis (as discussed in the next section). The final reason is to collect data on the nature of the evidence as it will be incorporated into the synthesis of evidence.

The type of information coded will depend upon the specific needs of a review. In different reviews the inclusion criteria will differ as will the information that is of interest in describing a research field. Similarly, quality appraisal will vary on the issues described in the next section. Data for the synthesis will be dependent on what findings are available from each study. Care will need to be taken to ensure that there not multiple findings from one study which result in over representation of that study in the synthesis. In meta-evaluations of mega-events the synthesis is likely to contain many different types of data so the coding system needs to be capable of accepting such heterogeneity. This makes it likely that coding will include both a priori closed categories and open coding of information (see Oliver and Sutcliffe 2012).

(iv) Stage 4: QUALITY AND RELEVANCE APPRAISAL

Step 4.1: Rigour by which the information has been produced

Step 4.2: Fitness for purpose of the method by which the information was produced for answering the review questions or sub-questions

Step 4.3: Fitness for purpose of the focus of the information (such as intervention, context, outcomes) for answering the review questions or sub-questions

Step 4.4: Overall weight of evidence that the information provides in answering the review questions or sub-questions

As already discussed, some forms of meta-evaluation are in themselves the application of standards to evaluate evaluations. This may be to develop formative feedback for a planned or ongoing study, an assessment of trustworthiness, an appraisal of the state of research, or a benchmark of quality standards. In meta-evaluations that synthesize the findings of evaluation studies there is a need to appraise the worth of studies to be part of that synthesis. If evaluations included in the Meta-Evaluation of the Olympic and Paralympic Games, for example, are not of good quality or relevant then the findings and conclusions of the Meta-Evaluation may not be valid.

The syntheses will be driven by questions that may be different from those considered by individual studies and so there is a need to interrogate these individual studies for results and process data that is relevant and trustworthy for answering the specific syntheses questions. The nature of quality appraisal will also be different for an aggregative review with pre specified methods (including quality appraisal) than a configuring review with more iterative concepts

and methods and emergent ideas about what is a good quality study in the review. This section considers some of the dimensions of quality appraisal in such syntheses (for further detail see Harden and Gough, 2012).

Dimensions of quality and relevance

The range of different purposes and dimensions of quality appraisal mean that there is a corresponding wide range of data that could be subjected to quality appraisal judgments and these data may be from any part or stage of the research study.

The standard dimension for assessing research is its quality in terms of creating knowledge, or epistemic value. For example, there may be agreed standards for executing certain research methods and those methods may be associated with achieving certain outcomes. Even if everyone agrees on these aims and methods, the reality is that an individual study may not follow these standards. There may be aspects of the method or its execution that deviate from these ideals. A study can thus be judged on how well it is executed according to agreed standards *and* the fitness for purpose of that method for answering the research question of the study. Furlong and Oancea (2008), however, argue for further dimensions such as applied use of the research (technical value), value of personal growth and engagement with users (capacity building and value for people) and cost-effectiveness and competitiveness (economic value).

A broad framework is provided by Pawson and colleagues (2003) who proposed the acronym TAPUPAS, for judging and interpreting the quality and usefulness of research and sources of evidence. This has six generic and one knowledge-specific dimension, and a set of indicative questions/statements against which each source can be appraised:

- **Transparency**. Is it open to scrutiny? Is it easy to tell how the evidence was generated?
- **Accuracy**. Is it well grounded? Are the recommendations and conclusions based on data or are they just asserted with little basis in the research itself?
- **Purposivity**. Is it fit for the purpose? Was the methodology a good fit for the types of questions being researched?
- **Utility**. Is it fit for use? Can information presented be used by others in the field, or is it incomplete or missing important information that would help in practical use?
- **Propriety**. Is it legal and ethical? Was the research conducted with the consent of stakeholders and within ethical guidelines?
- **Accessibility**. Is it intelligible? Is the information presented in a way that allows those who need it to readily understand and use it?
- **Specificity**. Whether the knowledge meets the specific standards that are already associated with that type of knowledge (e.g. practitioner, policy, research knowledge). Are there specific standards in a field that come into play?

Once the evidence has been judged (and possibly scored) against each of the criteria, an overall interpretation can be reached on quality. [9] This is similar to frameworks for assessing evidence according to evaluation standards, such as that created by the Joint Committee on Standards for Educational Evaluation in the United States (Yarbrough et al 2011). This incorporates over 20 evaluation standards across similar dimensions of: accountability, accuracy (covering research validity and aspects of purposivity), utility, propriety (including accessibility) and feasibility (covering such aspects as efficiency and viability)[10]. Whilst this framework has the more specific

[9] For more details and worked examples see: http://www.scie.org.uk/publications/knowledgereviews/kr03.pdf
[10] An outline of the standards and statements employed can be found at: http://www.jcsee.org/program-evaluation-standards/program-evaluation-standards-statements

aim of promoting quality in evaluation practice (and hence incorporates significantly more detail than is likely to be required by the Meta-Evaluation of the 2012 Olympic and Paralympic Games) some of the specific standard statements are relevant here.

Given the wide ranging purposes of quality appraisal associated with the Meta-Evaluation of the Olympic and Paralympic Games (combining elements of design, process and results meta-evaluation), a combination of generic criteria from existing frameworks such as that of Pawson et al (2004), and more specific meta-evaluation and even emergent criteria are likely to be required, and combined to produce a suitable appraisal tool.

In the 'systematic review of evidence' form of meta-evaluation, the evaluation of studies for inclusion in the synthesis depends on three main dimensions, including not only the technical quality of the evaluation, but also the fitness for purpose of that method for the review; and the validity of the approach used in the study relative to the review question (Gough, 2007). The concept of utility is particularly relevant here because, however, technically good a study is, it may not be fit for purpose for the meta-evaluation; the same study could be of high quality for one purpose but not for another (Stufflebeam 1981).

In sum, these distinctions represent three dimensions of: (A) technical adequacy of the execution of study; (B) relevance of the research design for the review question; and (C) relevance of execution of that design, which all combine to affect the weight of evidence that can be put on the study in answering the study's research question (Gough 2007). Weight of Evidence dimension A (WoE A) is a generic measure of how well a study has been executed within normal standards, whereas dimensions B and C (WoE B and C) are (meta) evaluation specific criteria.

In a synthesis of findings meta-evaluation, and in the case of the meta-evaluation of a mega-event, studies undertaken for many different reasons may be included in the synthesis. The weight of evidence system allows the meta-evaluator (reviewer) to assess the worth of the study in answering the review question, not necessarily the question of the authors of each piece of primary research. The evaluator of a study may also not share the perspectives of the author of the primary research and this is easily accounted for in the weight of evidence system.

In practice, these dimensions are applied in different ways. In terms of the first dimension of technical execution, there are a large number of checklists, scales or 'tools' available. Sometimes these can simply be prompts, as in this list from Dixon Woods and colleagues (2006) to help reviewers make judgements about the quality of papers within their review, which included a diverse range of study types on the topic of access to healthcare:

- Are the aims and objectives of the research clearly stated?
- Is the research design clearly specified and appropriate for the aims and objectives of the research?
- Do the researchers provide a clear account of the process by which their findings we reproduced?
- Do the researchers display enough data to support their interpretations and conclusions?
- Is the method of analysis appropriate and adequately explicated?

There are also some scales for different types of study design (a range of relatively short scales for different study designs can be found on the CASP website at http://www.casp-uk.net/). An example of a well known tool for evaluating impact studies for example is the Maryland Scale of Scientific Methods, which was developed to help identify what works in crime prevention, through ranking existing evaluations and studies from 1 (weakest) to 5 (strongest)

in terms of overall internal validity. Its implicit aims were also to encourage greater scientific rigour in future evaluations[11]. There are also many scales attempting to assess the adequacy of qualitative research. Spencer and colleagues (2003), for example, provide an array of measures, though the choice of the measures one might want to select in a particular review would depend upon the type of review questions being asked.

In terms of the dimension of the fitness for purpose of different research designs to the meta-evaluation, this is often in practice determined by the reviewer specifying in advance which types of research design will be included in the review (i.e. study design is part of the inclusion criteria). In other reviews, the reviewer makes a judgement as to the worth of the results of such a design (along with decisions on the other two dimensions of execution and validity) for answering the review question and thus the extent that the findings of the study will or will not be included in the synthesis.

In terms of the dimension of the focus of the study and the validity of the findings in terms of the review question, this is also often determined by inclusion criteria and by later reviewer judgements of adequacy. An example, relevant to sports mega-events, is the outcome measure of participation in sport. An outcome measure simply asking people if they intend to participate in sport may, for example, not be a valid measure of actual participation.

Judgement is also necessary for combining dimensions to make any overall conclusions on quality. A study may, for example, be strong in terms of study design but be poorly executed; a not quite so relevant but very well executed research design may provide more useful information to a synthesis. Similarly, a study may have a very appropriate design and be mainly well executed but use outcome measures that are not very valid for the synthesis. An example of this process is given in Appendix 2.

In addition, there is the issue of what decision is made on the basis of the evaluation of quality and relevance. Studies can be excluded, they can be tested as to their effect on the synthesis and excluded if this is out of line with other studies (test for sensitivity), they can be included but weighted in their contribution to the synthesis according to their quality/relevance, or the studies can all be included with their quality/relevance appraisal being provided to readers.

In sum, quality and relevance appraisal is based on methodological principles but there is variation in how these can be applied, so judgement is required with transparency within the meta-evaluation on the how decisions were made.

Critical appraisal of meta-evaluations

As well as evaluating studies included in meta-evaluations, the meta-evaluation as a whole can be critically appraised. This can be undertaken using any of the dimensions of appraisal discussed above. A particular area of potential poor meta-evaluation practice is a lack of independence of the meta-evaluators from the primary evaluations under study. If the people who are searching for, appraising and synthesizing studies were also the authors of some of the studies involved then these researchers may unwittingly be biased in their judgements and therefore in the results that they find. For the Meta-Evaluation of the Olympic and Paralympic Games, the individual studies are mostly being undertaken by others. Where members of the Meta-Evaluation team are involved in the generation of data for synthesis, this is to complete a missing part of the knowledge needed for the Meta-Evaluation, or else there is a strong separation from the evidence appraiser.

[11] See https://www.ncjrs.gov/pdffiles/171676.PDF for more details and the assessment framework employed.

5.3 Stage 5: SYNTHESIS

(i) Step 5.1: Clarify the evidence available for answering the review questions and sub-questions

(ii) Step 5.2: Examine patterns in the data and the evidence they provide in addressing review questions and sub-questions

(iii) Step 5.3: Combine sub-questions to address main questions and cross cutting themes

(iv) Step 5.4: Test the robustness of the syntheses

Synthesis is achieved by using the research questions to interrogate the available data to determine the weight of evidence confirmatory or in contradiction of all the component parts of the evaluative framework, and thus for answering all parts of the sub-questions and headline questions (Thomas et al 2012).

The main research questions drive a 'top down' approach to identifying sub-questions and relevant evidence. Yet the synthesis is largely achieved through a 'bottom up' approach, where evidence is combined to address more narrowly focused sub-questions, the answers to which are then themselves combined to address the more macro headline and cross-cutting questions.

Any sub-question for example may be addressed by a number of different types of data that explore any part of the hypothesized causative models and these elements of data may be analysed separately before being combined to address the sub-question. In effect therefore, synthesis is a process of multiple syntheses which may involve several parallel or hierarchical sub-syntheses within one sub-question, let alone the combination of several sub-questions to address headline questions.

Synthesis has a number of practical stages:

5.1) Clarify the evidence available for answering the review questions and sub-questions

First is clarification of the data available to be interrogated to answer the review questions. The specification of the questions and sub-questions and their evaluative and conceptual frameworks should already be clear as it is the starting point of the review process (though it may have undergone some iteration as the review has progressed and new sub-themes have emerged). The data to answer these questions will then have been identified from studies meeting the inclusion criteria during the data extraction phase, and will have been appraised as being of sufficient quality and relevance for either full inclusion or qualified inclusion in the synthesis. In the meta-evaluation of mega-events such as the Olympic and Paralympic Games, there will be a wide range of data to be considered including: focused evaluations of specific interventions in terms of specific outcomes; raw output data from different interventions; 'top down' national statistical/survey data; additional primary research to fill in missing data needs; and economic modelling of the impacts of the event.

5.2) Examine patterns in the data and the evidence they provide in addressing review questions and sub-questions

The review question is then used to drive the examination of patterns in the data. The review questions in this type of meta-evaluation are often driven by hypotheses of the role and impact of an intervention, in which case the patterns sought will be the ones related to the potential relationship between hypothesized independent and dependent variables.

This process may employ different methods of synthesis, depending upon the nature of the data and the evaluative framework. The inclusion criteria of the review questions and sub-questions may have limited the data to those of a similar type and allow an aggregated view of

the data. For example where the data is numerical then it may be possible to aggregate data statistically (as in the statistical meta-analysis of the results of experimental trials). Where this is not possible, due to lack of appropriate statistical data, then the synthesis may be limited to thematic summaries structured around the hypotheses being tested. If all of the data is of high quality and points in one direction, confirming or disconfirming the hypothesis, then it is nonetheless easier to justify conclusions. If the results are mixed then it is difficult to draw firm conclusions. More generally, counting up studies with different results in order to provide an overall judgement in relation to a hypothesis can be very misleading as the studies may be of differential power, quality and contextual relevance.

When the data is very varied the process of seeking patterns may require mixed methods approaches. These consider the relative internal and external validity, transferability of such qualified data, and the potential for triangulation of the data to enable confirmation or explanation (Teddie and Tashakkori 2009). This can entail using different types of data at one time to answer a single question. Alternatively, it may involve splitting the data into types and interrogating this separately in parallel before combining the results together to answer the review question.

In all of these approaches, it is the question (and sub-questions) that are driving the seeking of patterns and the methods of interrogation of the data. The detail of the questions and their conceptual framework, for example an explicit theory of change, drives the process.

5.3) Combine sub-questions to address main questions and cross cutting themes

Synthesis may also involve sub-component syntheses where different aspects of an issue have been interrogated by sub-questions (for example testing out confirmatory or contradictory explanations). These may include testing similar hypotheses or may involve checking some other specific part of a causative model; for example the prevalence of necessary preconditions. The way that the patterns are analysed again depends upon the evaluative framework and the specific methods of review chosen. This may include mixed data and thus mixed methods analysis.

Although linking together sub-questions is complex, the process is essentially no different from mixed methods analysis undertaken within one question. Again, this process may be undertaken by directly examining and combining the data related to each question or by doing this separately in parallel and then combining the results of the sub-sections.

5.4) Test the robustness of the syntheses

This involves testing the robustness of the syntheses by taking a critical examination of the extent that they appropriately use available data to answer the meta-evaluation questions. This may include providing qualifications to any conclusions, due for example to a lack of appropriate data to provide clearer answers to the initial meta-evaluation question.

This may also involve consultation with various stakeholders to ask about their interpretation, understanding and agreement with the interrogation and interpretation of the data; the interpretation of the data may vary between stakeholders, just as their initial questions and value-interests may vary. It is important that this is reflected in the approach to conducting and presenting the synthesis.

In sum, there are very many different types of review questions or sub questions that can be asked and many different synthesis techniques that can be applied. Synthesis is thus not a simple stage of review but a complex process that brings together the original question, the data available and different stakeholder judgements to attempt to answer each question.

5.4 Stage 6: CONCLUSIONS AND DISSEMINATION

(i) Step 6.1: Engage with users of the meta-evaluation to interpret draft findings

(ii) Step 6.2: Interpret and test findings

(iii) Step 6.3: Assess strengths of the review

(iv) Step 6.4: Assess limitations of the review

(v) Step 6.5: Conclude what answers can be given to questions and sub-questions from evidence identified

(vi) Step 6.6: Refine theories in light of evidence

(vii) Step 6.7: Disseminate findings

As a meta-evaluation is being undertaken for particular purposes, then those determining those purposes have a role in defining the questions, the evaluative framework and the interpretation of the results of the meta-evaluation. This should not create hidden bias. On the contrary, it should make explicit and consistent the perspectives (and values) driving the meta-level analysis of evidence and its judgements. The process of interpretation may include the reality testing of the results to check their relevance for different contexts, and from multi-stakeholder perspectives. Ideally, evaluation findings should also be reported in such a way that stakeholders can form their own judgements about those elements of the evaluation where they interpret the data differently.

The resulting conclusions also need to be presented in terms of the strengths and weaknesses of the meta-evaluation that produced them, in terms of the extent that the research was appropriately formulated and executed and reported. Overall, any study will be weakened if the problem has not been properly formulated, if inappropriate methods are selected to address that problem, if there is poor execution of methods (however appropriate they may be), and if the reporting is not clear so that it not be possible to appraise whether the evaluation was fit for purpose in method or undertaken correctly. For the Meta-Evaluation of the Olympic and Paralympic Games there are issues of quality appraisal for each of the stages of the work. This is complex when there are many themes, overall questions and sub-questions, and when many different forms of data are being used to address each of these question points.

Once the results of a meta-evaluation have been interpreted, tested and qualified they can be reported to others. In order to ensure transparency and accountability, this needs to include a full account of the methods of the meta-evaluation and the rationale for decisions taken. In order to ensure impact, this also requires methods to ensure the visibility, understandability, relevance and thus communication of the meta-evaluation conclusions.

Appendix 1: References

Ashworth K, Cebulla A, Greenberg D, Walker R. Meta-Evaluation: Discovering What Works Best in Welfare Provision. *Evaluation* 2004 10: 193

Barnett-Page E, Thomas J (2009) Methods for the synthesis of qualitative research: a critical review. *BMC Medical Research Methodology*, 9:59. doi:10.1186/1471-2288-9-59

Bickman, L. (1997). Evaluating evaluation: Where do we go from here? *Evaluation Practice,* 18, 1-16.

Bollen, K, Paxton, P, and Morishima, R (2005), 'Assessing international evaluations: An example from

USAID's gemocracy and Governance program', American Journal of Evaluation, 26 (2), 189-203.

Bornmann L, Mittag S, Daniel HD, (2006) Quality assurance in higher education – meta-evaluation of multi-stage evaluation procedures in Germany. *Higher Education* 52: 687–709

Bornmann L, Leydesdorff L, Van den Besselaar P. (2010). A meta-evaluation of scientific research proposals: Different ways of comparing rejected to awarded applications. *Journal of Informetrics*. Vol.4 No.3 pp211

Britten, N., Campbell, R., Pope, C., Donovan, J., Morgan, M., Pill, R. (2002) Synthesis of qualitative research: a worked example using meta ethnography. *Journal of Health Services Research and Policy* 7: 209-16.

Bustelo M (2003a) Evaluation of Gender Mainstreaming. Ideas from a Meta-Evaluation Study. *Evaluation* vol. 9 no. 4 383-403

Bustelo, M. (2003b) Metaevaluation as a tool for the improvement and development of

the evaluation functions in public administrations. Accessed 30th May 2011 at: http://cjpe.ca/distribution/20021010_bustelo_maria.pdf

Caird J, Rees R, Kavanagh J, Sutcliffe K, Oliver K, Dickson K, Woodman J, Barnett-Page E, Thomas J (2010) *The socioeconomic value of nursing and midwifery: a rapid systematic review of reviews.* London: EPPI Centre, Social Science Research Unit, Institute of Education, University of London.

Cook TD, Gruder C L (1978). Metaevaluation research. *Evaluation Review*, 2(1), 5-51.

Cooksy, L. J., & Caracelli, V. J. (2005) Quality, Context, and Use Issues in Achieving the Goals of Metaevaluation. *American Journal of Evaluation* 2005; 26; 31

Cooksy, L. J., & Caracelli, V. J. (2009a). Metaevaluation in practice: Selection and application of criteria. *Journal of MultiDisciplinary Evaluation*, 6 (11).

Curran V, Christopher J, Lemire F, Collins A, Barrett B (2003) Application of a responsive evaluation approach in medical education. *Medical Education* 37:256–266

DCMS (2011a) Report 1: Scope, research questions and data strategy Meta-Evaluation of the Impacts and Legacy of the London 2012 Olympic Games and Paralympic Games. London: Department for Culture, Media and Sport.

DCMS (2011b) Report 2 – Methods: Meta-Evaluation of the Impacts and Legacy of the London 2012 Olympic Games and Paralympic Games. London: Department for Culture, Media and Sport.

Dixon-Woods M, Cavers D, Agarwal S, Annandale E, Arthur A, Harvey J, Hsu R, Katbamna S, Olsen R, Smith L, Riley R, Sutton AJ (2006) Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology*. 6(35)

Eureval (2008) Meta-study on decentralised agencies: cross-cutting analysis of evaluation findings

Final Report September 2008 Evaluation for the European Commission Contract ABAC-101930. Accessed 30th May 2011 at: http://ec.europa.eu/dgs/secretariat_general/evaluation/docs/study_decentralised_agencies_en.pdf

Furlong J, Oancea A (2005) *Assessing Quality in Applied and Practice-based Educational Research: A framework for discussion.* Oxford: Oxford University Department of Educational Studies.

Garcia, B., Melville, R. and Cox, T. (2010) *Creating an impact: Liverpool's experience as a European capital of culture*. Liverpool: University of Liverpool.

Gough D (2007) Weight of evidence: a framework for the appraisal of the quality and relevance of evidence In J. Furlong, A. Oancea (Eds.) Applied and Practice-based Research. Special Edition of Research Papers in Education, 22, (2), 213-228.

Gough D, Thomas J (2012). Commonality and diversity in reviews. In Gough, D et al. *Introduction to systematic reviews*. London: Sage.

Gough D, Thomas J. Oliver S (2012) Clarifying differences between systematic reviews. Systematic Reviews Journal. (1:28)

Green, J.C., Dumont, J., & Doughty, J. (1992) A formative audit of the ECAETC year 1 study evaluation: Audit procedures, findings, and issues. *Evaluation and Program Planning, 15*(1), 81-90.

Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O, Peacock R (2005a) Storylines of research in diffusion of innovation: a meta-narrative approach to systematic review. *Social Science & Medicine*. 61: 417–430.

Hanssen CE, Lawrenz F,Dunet DO, (2008) Concurrent Meta-Evaluation: A Critique. *American Journal of Evaluation*. Volume 29 Number 4 December 2008 572-582

Harden A, Gough D (2012) Quality and relevance appraisal. In Gough, D et al. Introduction *to systematic reviews*. London: Sage.

Langen, F. and Garcia, B. (2009) *Measuring the impacts of large scale social events: a literature review*. Liverpool: John Moores University.

London Assembly (2007) *A Lasting Legacy for London? Assessing the legacy of the Olympic Games and Paralympic Games*. London: University of East London.

Madzivhandila TP, Griffith GR , Fleming E, Nesamvuni AE (2010) *Meta-evaluations in government and government institutions: A case study example from the Australian Centre for International Agricultural Research*. Paper presented at the Annual Conference Australian Agricultural and Resource Economics Society (AARES), 8-12 February 2010. Accessed 30th May 2011 at: http://ageconsearch.umn.edu/bitstream/59098/2/Madzivhandila,%20Percy.pdf

Martin, Downe and Bovaird (2012), *Policy and Politics* (forthcoming).

Maxwell GS (1984) A Rating Scale for Assessing the Quality of Responsive/Illuminative Evaluations. *Educational Evaluation And Policy Analysis.* 6; 131

Noblit GW, Hare RD: *Meta-Ethnography: Synthesizing Qualitative Studies*. London: Sage; 1988.

Oliver S, Bagnall A M, Thomas J, Shepherd J, Sowden A, White I, *et al*. Randomised controlled trials for policy interventions: a review of reviews and meta-regression. *Health Technol Assess* 2010;**14**(16).

Pawson R (2002) *Does Megan's Law Work: A theory-driven systematic review*. ESRC UK Centre for Evidence Based Policy and Practice, Working Paper No 8 (available at

www.evidencenetwork.org).

Pawson R, Grrenhalgh T, Harvey G, Walshe K (2004) Realist synthesis: an introduction. ESRC Research Methods Programme., University of Manchester. RMP Methods Paper 2/2004

Pawson, R. (2006) *Evidence Based Policy: A Realist Perspective,* London: Sage.

Pawson R, Boaz A, Grayson L, Long A, Barnes C (2003) *Types and Quality of Knowledge in Social Care*. London: Social Care Institute for Excellence.

Petrosino A, Turpin-Petrosino C, Buehler J. "Scared Straight" and other juvenile awareness programs for preventing juvenile delinquency. *Cochrane Database of Systematic Reviews* 2002, Issue 2. Art. No.: CD002796. DOI: 10.1002/14651858.CD002796

Rand Europe (undated). Setting the Agenda for an evidence-based Olympics Setting the evidence-based agenda: a meta-analysis, RAND Europe. (1942).

Rodgers M, Sowden A, Petticrew M, Arai L, Roberts H, Britten N, Popay J. Testing methodological guidance on the conduct of narrative synthesis in systematic reviews: effectiveness of interventions to promote smoke alarm ownership and function. *Evaluation*. 2009 ; 15 (1): 49-73

St Pierre R (1982) Follow Through: A Case Study in Meta-Evaluation Research

Educational Evaluation and Policy Analysis. Vol. 4, No. 1 (Spring), pp. 47-55

Sandelowski M. Voils CJ, Leeman J, Crandlee JL. (2011)Mapping the Mixed Methods-

Mixed Research Synthesis Terrain. *Journal of Mixed Methods Research.* Published on-line 2011

Shaffer, M. Greer, A. and Mauboules, C. (2003) *Olympic Costs & Benefits A Cost-Benefit Analysis of the Proposed Vancouver 2010 Winter Olympic and Paralympic Games*. Candian Center for Policy Alternative: British Columbia.

Schwandt (1992) constructing appropriate and useful metaevaluative frameworks. *Evaluation and Progam Planning*, 15, 95-100.

Schwartz R, Mayne J (2005) Assuring the quality of evaluative information: theory and practice. *Evaluation and Program Planning*. 28, 1–14

Scriven M (1969). An Introduction to meta-evaluation, *Educational Products Report*, 2, 36-38.

Scriven M (1998) *The Nature of Evaluation. Part I: Relation to Psychology*. ERIC/AE Digest. Washington DC: ERIC Clearinghouse on Assessment and Evaluation.

Shadish W (1998) *Some Evaluation Questions*. ERIC/AE Digest. College Park: ERIC.

Smith, M. (2008) *When the Games Come to Town: Host Cities and the Local Impacts of the Olympics: A report on the impacts of the Olympic Games and Paralympics on host cities*, London East Research Institute Working Paper.

Spencer L, Ritchie J, Lewis J, Dillon L(2003) Quality in Qualitative Evaluation: A framework for assessing research evidence. London: National Centre for Social Research

Stufflebeam, DL. (1981). Metaevaluation: Concepts, standards, and uses. In R.A. Berk (Ed.), *Educational evaluation methodology: The state of the art* (pp. 146-63). Baltimore, MD: Johns Hopkins University Press.

Stufflebeam, DL. (2001). The Metaevaluation Imperative. *American Journal of Evaluation* 2001 22: 183

Tashakkori A, Teddlie C (eds.) Handbook of Mixed Methods in the Social and Behavioral Sciences (2nd edition). New York: Sage

Thomas J, Harden A, Newman M (2012) Synthesis: combining results systematically and appropriately, . In Gough, D et al. *Introduction to systematic reviews.* London: Sage.

Thomas J, Harden A, Oakley A, Oliver S, Sutcliffe K, Rees R, Brunton G, Kavanagh J (2004) Integrating qualitative research with trials in systematic reviews: an example from public health. *British Medical Journal*. 328: 1010-1012.

Thomas MR (1984) Mapping Meta-Territory. *Educational Researcher*. 13; 16

Torgerson CJ, Gorard S, Low G, Ainsworth H, See BH, Wright K (2008) What are the factors that promote high post-16 participation of many minority ethnic groups? A focused review of the UK-based aspirations literature. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Vigor, A. Mean, M. and Tims, C. (2004) *After the Goldrush: a sustainable Olympics for London* (IPPR and Demos: London).

Voils, C. I., Sandelowski, M., Barroso, J., & Hasselblad, V. (2008). Making sense of qualitative and quantitative research findings in mixed research synthesis studies. *Field Methods,* 20, 3-25.Warwick I, Mooney A, Oliver C (2009) *National Healthy Schools Programme: Developing the Evidence Base.* Project Report. Thomas Coram Research Unit, Institute of Education, University of London, London.

Wingate LA (2009) Meta-evaluation: Purpose, Prescription, and Practice

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., and Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users* (3rd ed.). Thousand Oaks, CA: Sage.

Appendix 2: Example of weight of evidence coding

Review question: What are the factors that promote high post-16 participation of many minority ethnic groups?

Review overview: The desire to widen participation in formal post-compulsory education and training is a policy agenda common to most developed countries. Given that some minority ethnic groups have higher rates of post 16 participation in the UK than both the majority white cohort and some other minorities, identifying potential determinants could lead to a method of increasing participation for all. The aim of this review, therefore, was to determine the factors that drive high post-16 participation of many minority ethnic groups. Studies had to be conducted in the UK, have a key focus on post-16 aspirations, provide a distinct analysis of different minority ethnic groups and either a) elicit student aspirations about education (cross-sectional survey or qualitative study) or b) investigate the statistical relationship between aspirations and educational variables (secondary data analysis). A conceptual framework for the synthesis was constructed to capture post-16 'promoters' and 'non-promoters' within the following categories: government policy; institutional practices; external agencies; work; religion; family; individual aspirations; and other factors.

Weight of Evidence (WoE): Separate ways of assessing studies were put in place for the two different types of studies included in the review. For all dimensions of WoE, studies were given a rating of low, medium or high. Examples of how studies were judged 'high' or 'medium' are shown below. A standard formula was used to calculate the overall weight of evidence for a study (e.g. for a study to be rated overall 'high', it had to be rated 'high' for WoE A and B and at least 'medium' for WoE C). Only the findings from studies rated 'high' or 'medium' were used in the synthesis stage of the review.

	Cross-sectional surveys and qualitative research	**Secondary data analysis**
WoE A: Soundness of studies	High: Explicit and detailed methods and results sections for data collection and analysis; interpretation clearly warranted from findings Medium: Satisfactory methods and results sections for data collection and analysis; interpretation partially warranted from findings.	High: Explicit and detailed methods and results sections for data analysis; interpretation clearly warranted from findings. Medium: Satisfactory methods and results sections for data analysis; interpretation partially warranted from findings.
WoE B: Appropriateness of study design for answering the review question	High: Large scale survey methods using questionnaires and/or interviews. Medium: Survey methods using questionnaires and/or interviews.	High: Large scale secondary data analysis; origin of dataset clearly stated Medium: Secondary data analysis; origin of data set partially indicated
WoE C: Relevance of the study focus to the review	High: Large sample, with diverse ethnic groups, with good generalisability and clear post-16 focus. Medium: Adequate sample, with diverse ethnic groups, with generalisability and partial post-16 focus.	High: Large sample, with diverse ethnic groups, with good generalisability and clear post-16 focus, and low attrition from original dataset Medium: Adequate sample, with diverse ethnic groups, with generalisability and partial post-16 focus, and any attrition indicated

NB: Additional guidance was provided for reviewers for making judgements (e.g. what constitutes a large sample.

Source: Adapted from Torgeson et al 2008.

Appendix 3: Expert interviewees

Professor Maurice Basle, University of Rennes, France

Professor Tony Bovaird, Birmingham University, UK

Professor Ann Buchanan, University of Oxford, UK

Professor Tom Cook, North Western University, US

Professor Henri de Groot, Vrije Universiteit, Netherlands

Professor Michael Hughes, Audit Commission, UK

Professor Paul Lawless, Sheffield Hallam University, UK

Luc Lefebvre, SEE, Luxembourg and Institut d'Etudes Politiques, Paris

Michiel de Nooij, SEO Economisch Onderzoek, Netherlands

Audrey Pendleton, Institute of Education Services, US

Professor Elliot Stern, Bristol University, UK

Daniela Stoicescu, Ecorys

Jacques Toulemonde, Eureval, France

Appendix 4: Interview Topic Guide

Introduction

- Thank interviewee for participating.
- Explain purpose of the study and focus of interview.
- Ask for permission to list interviewee's name in report and to quote them on non attributable basis.

Definitions

- What involvement have you had in meta-evaluation?
  (Prompt – as an evaluator, practitioner, commentator, peer reviewer etc)
- What do you understand by the term 'meta-evaluation'?
- In your view what differentiates meta-evaluation from other forms of evaluation?
  (Possible prompts – purpose, methods, uses to which it is put)

State of the art

- How would you describe the current state of the art of meta-evaluation?
  (Possible prompts – good, patchy, confused, underdeveloped etc.)
- What do you see as the main strengths (if any) of current meta-evaluation practice?
- What do you see as the main gaps/weaknesses (if any)?

Methodologies

- Based on your own experience what do you see as the main methodological challenges for meta-evaluation?
- What (if anything) do you think can be done to improve meta-evaluation methods?
- (Possible prompts – what new methods and approaches are required?)
- Are there any approaches that you think work particularly well in integrating the findings of separate studies? (Prompt - as in the case of the 2012 Olympic and Paralympic Games?)
- Would you recommend any particular approaches to identifying interdependencies between the different kinds of outcomes associated with complex interventions?
  (Prompt - for example economic, social and sport outcomes)

Exemplars and interviewees

- Are there any specific examples of meta-evaluations which you'd recommend we look at (including any of your own work)? (Possible prompts: Good practice, mega-events, innovative methods)
- Could you describe briefly the methods which were used in this study?
- In your view how successful was the meta-evaluation, and what were the key methodological lessons?

- Can we access the final report and key research tools from the study?
- Is there anyone else who you think we should talk with about these issues?

Follow up

- Would you be interesting in staying in touch with the study? If yes what kind of involvement would you consider?

THANK INTERVIEWEE AND EXPLAIN NEXT STEPS IN STUDY