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

ESRC Methods Paper

Final Report – 10 March 2014

David Gough, Steve Martin, Tony Bovaird and Jonathan France

The authors of this report are:

David Gough, EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Steve Martin, Cardiff Business School.
Tony Bovaird, Institute of Local Government Studies; and Third Sector Research Centre, University of Birmingham.
Jonathan France, Ecorys

Please cite as:

Gough D, Martin S, Bovaird T, France J (2014) Meta-evaluation of the Impact and Legacy of the London 2012 Olympic Games and Paralympic Games: ESRC Methods Paper. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London.

ISBN: 978-1-907345-89-0

Contents

Executive Summary 1

Contents 1

1: Introduction 15

2: Literature on mega-events 19

3: Literature on meta-evaluation 24

4: Analysis of expert interviews 40

5: Guidelines for meta-evaluation 53

6: Learning from the Evaluation Team 68

7: Value to policy makers: lessons from the London 2012 meta-evaluation 78

Appendices 85

Executive Summary

Introduction

This report brings together the findings from the Developing Meta-Evaluation Methods study, which was undertaken in conjunction with the Meta-Evaluation of the Impacts and Legacy of the London 2012 Olympic and Paralympic Games (the Meta-Evaluation). The Meta-Evaluation was commissioned by the Department of Culture, Media and Sport (DCMS). The work on methods is funded by the Economic and Social Research Council (ESRC). The aim is to review and advance understanding of methods of meta-evaluation through addressing the following questions:

(i) How can we better define and conceptualize meta-evaluation/analysis?
(Sections 2 and 3 of this report)

(ii) What are the lessons from conducting previous meta-evaluations (at home and internationally) and how can meta-evaluation be improved?
(Section 4 of this report)

(iii) 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)?
(Section 5 of this report)

(iv) 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?
(Sections 6 and 7 of this report)

The project was a research and development exercise. The methods included reviews of the literature, interviews with experts, round tables and workshops with the Meta-Evaluation team and with policymakers.

The research literature on mega-events

Mega events

The Olympic and Paralympic Games represent a large mega-event with very many sub-components, the effects of which need to be studied together to provide an overall evaluation of impact. A brief review of the literature on the evaluation of a range of types of mega events identified many kinds of impacts. Almost all studies cover economic, social, and environmental impacts. Key themes or indicators have included improvements in reputation management, employment and skills, social capital, inclusion and well-being, environmental sustainability and governance capacity. The type of outcomes of interest to the Olympic Games have also varied over time, from a concern for enabling peace and understanding, through to economic impacts, and more recently to sustainability and securing longer-term legacy.

Evaluation of the impact of mega events is challenging as: they often have multiple objectives; their stated objectives evolve over time; different groups articulate different kinds of objectives; the objectives may be direct or indirect; and outcomes may be negative and/or unanticipated. Studies typically analyse each key objective or legacy theme or indicator separately, frequently including a chapter on each major theme of impact with a number of sub-themes. The studies often differentiate the kinds of impact measures and/or evaluation activity for different phases of the event from planning to delivery to legacy.

Different kinds of mega-event impacts and legacies require different measures and evaluation methodologies, and the mega events literature does not provide a grand conceptual amalgam capable of reflecting all ambitions. Outcomes can also be difficult to measure and the research designs are normally natural experiments with little experimental control or clear comparison groups, making any judgments of causal attribution imprecise. There is sometimes a sense that mega-events often leave some sort of overall lasting 'impression', but this is difficult to specify.

The research literature on meta-evaluation

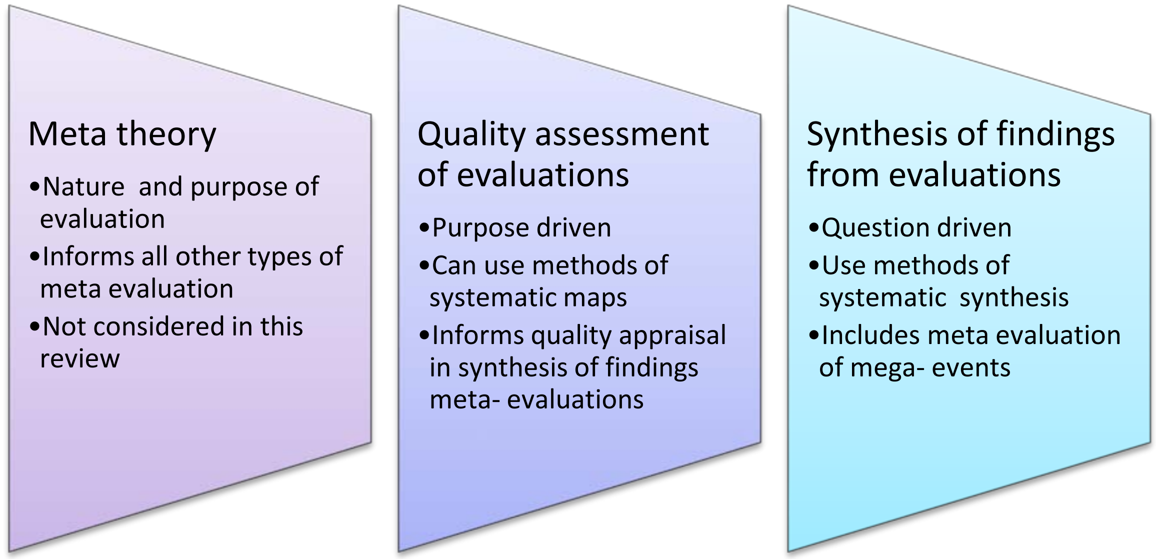
Meta-evaluation

Meta-evaluation is the 'evaluation of evaluations'. As mega-events consist of many sub-events and activities and their evaluation is of these combined components, evaluations of mega-events can be considered as one form of meta-evaluation. However beyond this the evaluation field does not seem to have a common agreement about meta-evaluation practice. The literature is also rich on conceptual issues, but thin on technical issues.

A brief review of this literature reveals wide variation in purpose and methods. Three main types of meta-evaluation can be identified: (i) the analysis (evaluation) of the nature and purpose of evaluation; (ii) the evaluation of the quality of evaluation studies; and (iii) the synthesis of the findings of individual studies to answer an overall evaluation research question.

All three forms of meta-evaluation have relevance for the evaluation of mega-events. Meta-theory raises fundamental issues about the nature and purpose of evaluations and is the building block for evaluation science. Quality assessment of evaluations can help develop good practice in both methodological and policy terms; it raises 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 aims to deliver a fuller understanding of the effectiveness of a policy initiative or programme (such as a mega-event).

Three Main Types of Meta-Evaluation


Meta theory
•Nature and purpose of evaluation
•Informs all other types of meta evaluation
•Not considered in this review
Quality assessment of evaluations
•Purpose driven
•Can use methods of systematic maps
•Informs quality appraisal in synthesis of findings meta- evaluations
Synthesis of findings from evaluations
•Question driven
•Use methods of systematic synthesis
•Includes meta evaluation of mega- events

Given the task in hand to support the Meta-Evaluation of the Impacts and Legacy of the 2012 Olympic and Paralympic Games, this report is most concerned with the synthesis of findings from existing evaluations (meta-evaluation Type iii), though some consideration is also given to the quality assessment of evaluations (Type ii).

Quality assessment meta-evaluation
The science of meta-evaluation in terms of assessing the quality of evaluations considers issues such the usefulness of a study, the assessment of a research team or organization (including researcher independence), and assessment of the strengths and weaknesses of a method. This has led to the creation of methodological standards for evaluation, which list criteria for evaluations and meta-evaluations relating to such aspects as utility, feasibility, propriety, accuracy and accountability.

Based upon the literature, there are different approaches to quality assessment meta-evaluations. Key dimensions on which they vary include: **aims** (which may include (i) evaluating the quality of a study to determine its trustworthiness; (ii) a broader remit of auditing the quality of studies and enabling the improvement of their quality; or (iii) the development of quality standards to make such trustworthiness assessments); and the evaluation **phase** in focus (which could be (i) the design of a study; (ii) the process by which a study is undertaken; or (iii) the results of a study). The **timing** of meta-evaluation may be: (i) concurrent and formative; or (ii) after the evaluation and summative.

In the case of the London 2012 Olympic and Paralympic Games mega-event, this meta-evaluation incorporates both formative and summative stages of assessment, and multiple purposes, including:

- 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);
- 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.

Synthesis meta-evaluation
Synthesis meta-evaluation can be considered as a particular kind of systematic review, which has its own detailed methods literature. Systematic reviews use transparent and rigorous methods to combine (aggregate) or arrange (configure) the findings of individual studies and their measurements or concepts to create an overall summary finding across studies. In a systematic review of impact evaluation studies, the findings of a number of individual evaluation studies are brought together into one large evaluation. This is a form of secondary research and requires specification of the research questions (and its assumptions), and methods of identification, appraisal, selection and synthesis of study findings to answer the review question. Types of review approaches include:

- Experimental assessments of the efficacy of an intervention
- Testing of causal theories
- Conceptualizing experience, meaning and process: configuring conceptual synthesis
- Complicated and complex mixed-methods systematic reviews
- Reviews of reviews
- Non systematic reviews of broad research questions

In general, reviews reflect the variation in **approaches and methods** found in primary research including the **research paradigm** and underlying epistemology. Aggregative reviews tend to be theory testing, use pre-specified concepts and methods and seek homogeneity of data. They can be 'black box' theory-light assessments of impact on specific outcomes or more theory driven and complex evaluations. Configuring reviews tend to generate or explore theory using iterative concepts and methods and seek heterogeneity. Systematic reviews can, depending on the research question and design, therefore be quantitative or qualitative in nature, or a mixture of both.

Reviews also vary in their **structure** and 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).

In terms of **detail**, 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 deeper the detail, the greater the challenge to manage the diversity of issues (as is likely to be 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 in order to be undertaken systematically with such little resource, or else lack **rigor** in method (for example a non-systematic scoping review).

Relevance of systematic review methods to the meta-evaluation of mega events
The synthesis of findings of evaluations often includes studies of separate examples of the same event or situation (for example a systematic review of many different studies evaluating the effectiveness of an intervention applied in similar but not exactly the same contexts), to deepen knowledge of results and/or process lessons. In the meta-evaluation of a mega-event it is relevant evaluation studies from different (but nonetheless related) sub-components of the same event that are brought together and synthesized, in order to provide a fuller understanding of the outcomes of the mega-event and its legacy. This implies additional layers of complexity; for example simultaneous sub-component meta-evaluations at various levels must first be carried out before the results of these thematic syntheses are combined to answer overarching questions about the impacts and lessons from mega events.

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 review and synthesis methods that need to be combined to address one or more overarching questions. If an overarching categorisation is sought, then the meta-evaluation of a mega-event might be considered to be closest in its aims and approach to a **mixed-methods aggregative theory testing review**.

Lessons from previous meta-evaluations

In addition to examining the research literature, the perspectives of recent practitioners of meta-evaluation were considered to help incorporate transferable lessons and evidence of what works from the research community. Interviews were conducted with **13 leading evaluation experts** working in academia or consultancy in the US and Europe.

Definitions
Like the broader evaluation community, the experts defined meta-evaluation in two main ways. Experts from the US defined it as an activity which **sets standards for evaluation activity and judges the quality of evaluations**. In the US, this is a well-established and well regarded activity that uses clearly defined criteria and methodologies to build capacity by educating researchers and research users about good evaluation practice. Traditionally, the focus has been on the design phase. But there has been a growing emphasis on also assessing the way evaluations are executed.

Most of the European experts defined meta-evaluation as the **synthesising of secondary data to provide an overall assessment of a series of related policies or interventions**. This also involves assessing the quality of the data that are used, but the distinctive feature of meta-evaluation here is the attempt to identify 'high level' outcomes and interactions between policies. The experts drew a distinction between meta-evaluation, systematic review and meta-analysis. All three activities seek to synthesise evidence from diverse studies or sources. But in their view, systematic reviews and meta-analyses are more narrowly focused than meta-evaluation and their methodologies are more clearly defined. Meta-evaluation methods are more varied and less well codified.

The State of the Art

The experts told us that though they are important, **synthesis meta-evaluations are rare, the literature about them is sparse and methods are underdeveloped**. Methods are borrowed from other branches of evaluation research and from across the social sciences more generally. However, studies are usually designed from scratch with little or no reference to previous meta-evaluations. So there is a **need to raise awareness of meta-evaluation and provide training in meta-evaluation methods.**

Politics of Meta-evaluation

The experts reported that because meta-evaluations study high-profile interventions, **politicians care about their results**. This enhances the prospects of utilisation but can prove problematic if they show that interventions have failed to produce the impacts which policy makers hoped for.

Data

According to the experts, the use of secondary data is a defining characteristic of synthesis meta-evaluations. But **in practice it is difficult to synthesise data collected at different times, by different teams, and for different purposes**. Ideally therefore, meta-evaluations should be commissioned first and meta-evaluators should be involved in the design of the studies that their work will draw on. Also, the scale and complexity of the interventions that meta-evaluations study mean that it can be **difficult to establish attribution because a wide range of factors can influence the observed outcomes**. The experts advised that it is, therefore, important to **focus meta-evaluations on a limited number of key themes** for which it is possible to **specify verifiable cause-and-effect mechanisms** that link interventions to outcomes. Where cause-and-effect mechanisms are not likely to be identifiable, or are not appropriate (e.g. where it is believe that the phenomena studied make up a complex adaptive system), then the meta-evaluation should point this out clearly, as it is essential that policy makers are aware of this and are not misled to trust misleading claims for the consequences of particular interventions.

A framework for conducting impact meta-evaluation

Although the practitioners interviewed differentiated systematic review from meta-evaluation, there are logical reasons for considering synthesis meta-evaluations to be a specialised form of systematic review; i.e. a specialized form of bringing together data from different studies to provide an overall assessment; particularly if undertaken robustly. The wide variety of impacts associated with a mega-event such as the Olympics however means that the specific type of data sought, the appraisal criteria deployed and the methods for synthesising data will differ widely across such meta-evaluations. It is nonetheless possible to provide generic guidance on how to structure such impact meta-evaluations, informed by systematic review methods. The structure presented below informed the particular steps and strategies undertaken as part of the Meta-Evaluation of the 2012 Games.

Stage 1: Defining the scope of the meta-evaluation

Step 1.1: Identify the scope 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 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 and systematic reviews highlights the importance of clarity about the research questions being addressed, the direct and indirect indicators to be used to address these questions, and the time span over which the questions are to be considered. The questions can be multiple and complex and at many different levels of analysis and of more or less concern to different stakeholders. Different individuals and groups will have different interests and thus different questions. Questions will contain theoretical and ideological assumptions of various types.

The research question asked by an impact meta-evaluation is essentially the testing of a hypothesis of a 'theory of change' (or multiple sub-theories of change in the case of a mega-event). The starting idea is that the event will have some positive (and maybe some negative) effects. The overall questions asked by a meta-evaluation are 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, as detailed across these theories of change and/or detailed logic models. An intervention may also have differential effects if provided in different ways to different groups in different situations. There will also be interactions and crossovers between different themes of activity. In addition, there may be cross cutting issues across themes for which questions may be identified a priori or which arise iteratively as the data is examined and the empirical data is considered. Questions, sub-questions, and cross cutting themes thus form the basis of an evaluative framework for seeking, organizing and analysing data.

Stage 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 evidence that is being sought to answer the meta-evaluation questions and sub-questions 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 data that is available in practice is, of course, also limited by the studies available. In the case of the Meta-Evaluation of the 2012 Games, this included 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.

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 there are at least five reasons for recording information from each study: (i) to describe the study in general ways to keep track of the study through the process of the review; (ii) to provide information in order to assess whether the data meets the inclusion criteria for the meta-evaluation and thus should be included in it; (iii) to be able to describe (or 'map') the field of research evidence meeting the inclusion criteria; (iv) to provide information to enable the quality and relevance appraisal of each piece of evidence to check how fit for purpose it is for the synthesis; and (v) to collect data on the evidence as it will be incorporated into the synthesis. Data for the synthesis will be dependent on what findings are available from each study relating to the mega-event. The synthesis is likely to contain many different types of data so the coding system needs to be capable of accepting such heterogeneity. Also, the same piece of data may be used in different ways in different parts of the meta-evaluation structure.

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, and 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

The standard dimension for assessing research is its quality in terms of creating knowledge, or epistemic value, and there are many scales available to support such judgments. However the study may be technically well executed but may not be well suited to answer the meta-evaluation question (and sub-questions) of the mega-event in terms of design or relevance, and so judgments are required on these dimensions too. Finally, judgment is necessary for combining dimensions to make any overall conclusions on quality. Studies can be then be excluded, included but weighted in their contribution to the synthesis, or included with their quality/relevance appraisal being provided to readers. As well as evaluating studies included in meta-evaluations, the meta-evaluation as a whole can be critically appraised. This can be undertaken using the same dimensions of appraisal.

Stage 5: Synthesis

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

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

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

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' (either confirmatory or contradictory) for all of the component parts of the evaluative framework, and thus for answering all parts of the sub-questions and headline questions.

The main research questions previously drove a 'top down' approach to identifying sub-questions and relevant evidence for the evaluation of the mega-event. However 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.

In effect, synthesis is a process of multiple syntheses which may involve several parallel or hierarchical sub-syntheses within one sub-question, let al.one the combination of several sub-questions to address headline questions.

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.

Stage 6: Conclusions and dissemination

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

Step 6.2: Interpret and test findings

Step 6.3: Assess strengths of the review

Step 6.4: Assess limitations of the review

Step 6.5: Answer questions and sub-questions from evidence identified

Step 6.6: Refine theories in light of evidence

Step 6.7: Communicate findings

The conclusions need to be presented in terms of the strengths and weaknesses of the meta-evaluation that produced them, i.e. in terms of the extent to which the research was appropriately formulated and executed and reported. 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.

For transparency and accountability, there should be a full account of the methods of the meta-evaluation and the rationale for decisions taken. In order to facilitate impactful findings, this also requires methods to ensure the visibility, clarity, relevance and communication of the meta-evaluation conclusions. This is of particular relevance to a mega-event meta-evaluation, where there are multiple interested audiences, with different interests and expectations.

Practical lessons from undertaking the Meta-Evaluation of the Impacts and Legacy of London 2012 Olympic and Paralympic Games

The study explored and compared the perceptions of the usefulness of the meta-evaluation held by the research team which undertook it and the policy community on whose behalf it was undertaken. Interviews and workshops were completed with the research team members and a policy workshop was convened with all the relevant policy departments in Whitehall, followed-up by correspondence on key issues.

Overall value of meta-evaluation

The overall assessment of both policy makers and the research team is that, in spite of the considerable challenges involved, **meta-evaluation is seen as having an important role to play in enabling an overall assessment of the impacts and legacy of high profile, large scale interventions and mega-events** of the kind that cannot be provided by more narrowly focused evaluations of individual projects and programmes. The research team recommended that other countries staging mega-events (including future Olympic and Paralympic Games) should conduct meta-evaluations.

The research team pointed to a wide range of useful lessons for policy makers, both in relation to the specific staging of mega-events and broader imperatives, e.g. the experience of the Games Makers is very relevant to the understanding how the use and value of volunteers can be maximised in society. Policy makers emphasized that the Meta-Evaluation had provided them with value-added in showing how their activities connected to outputs and outcomes in other programmes, in ways they had not previously identified. At a general level, it had highlighted relationships between their core activities and wider variables such as sustainability, impacts on people with disabilities, the potential role of volunteering, international reputation etc. More narrowly, it had helped in identifying the effects of some of their activities, where in the past they have not been able to afford the kind of detailed surveys or analysis which the Games have enabled. As a consequence of both of these sets of insights, important gaps had also been identified in their data on the cost-effectiveness of their activities.

Encouraging policy focus – but staying flexible

One of the lessons from the Meta-Evaluation is that it is **important to focus on core themes, which provide a clear focus for a meta-evaluation, but to combine this with an awareness of issues which cut across themes.**

Both the research team and policy makers agreed that **defining research questions and logic models at the outset had been a useful means of focusing a study on the key impacts and legacies and how these were being achieved.** However, the research team emphasized that these should be **reviewed regularly during the course of a study to take account of emerging findings, changing objectives, constraints on data availability and unintended and unanticipated outcomes.** This is even more important if it becomes evident during the course of a meta-evaluation that some elements of the policy system are better modelled as complex adaptive systems than as systems with predictable outcomes (although this was not an issue which arose during the 2012 Games Meta-Evaluation).

Policy makers suggested that, as policy developed, it would probably have been better to have had some checkpoints at which adjustments could be made to the research questions. This was either because new questions had become policy-relevant over time, or because positive or negative unintended consequences were becoming evident and the meta-evaluation process itself was challenging initial assumptions. In practice, it was felt that such changes would probably only apply to a small minority of research questions in the Meta-Evaluation of the 2012 Games, since most had stood the test of time. However, one of the key benefits of having a meta-evaluation team in place earlier on is the potential for identifying and responding to such emerging research needs.

An issue on which there was some disagreement, both among policy makers and in the research team, was whether the Meta-Evaluation might usefully have started with a much smaller core of questions, to which extra questions could have been added to over time as they surfaced. This would have encouraged greater realism about what can be achieved even by an avowedly ‘overarching’ evaluation, which inevitably faces many pressures to be highly ambitious. On the other hand, some felt strongly that it is the comprehensive nature of a meta-evaluation which constitutes its greatest value-added, when compared to other types of evaluation.

Need for a clear policy customer

In order for this policy focus to be achieved, **there is a need for a powerful 'customer' for meta-evaluation who is interested in the overall picture it presents**. A problem for both the research team and the Steering Group of the 2012 Games Meta-Evaluation was that there was not initially a strong sense of ownership of the Games across different government departments (beyond DCMS), and others who did get involved were not always able to state clearly what the priorities were for their organisations. This lack of ownership was perhaps one of the reasons that the research questions were difficult to narrow down – without partners who are clear about what their priorities are, it is harder to prune questions which might later turn out to be important. In practice, beyond the general commitment to legacy, clues about individual government priorities only emerged slowly and over time. Consequently, the research team had to treat all the research questions equally.

Key role of meta-evaluation at policy formation stage

There is clearly still a divide between the policy side of government and the research community, so that the flow of information remains sporadic and imperfect in both directions. The policy makers at the Policy Workshop suggested that the desire for evidence is most intense in government during the policy formation period. This is in line with the argument of Michael Quinn Patton (1997) that 'process use' is more important than 'use of substantive findings', since findings tend to have a short half-life, whereas process use teaches policy makers a new way of thinking and learning. Consequently, **it is essential that the meta-evaluation is put in place at a very early stage, so that its logic and findings can influence policy as it emerges**.

Data aggregation and synthesis

Interviewees suggested that one of the key lessons from their work is that because, by definition, meta-evaluations will be dependent on other studies for evidence, meta-evaluators are not in control of the data available to them, the form it takes or when it becomes available. This means that there is a premium on synchronising the design of meta-evaluations and the studies they draw on. Some of the challenges around data availability can be addressed if **meta-evaluations are commissioned ahead of the studies which they will draw upon**. It is also **helpful if meta-evaluators have an input into the terms of reference of the studies on which they will draw and on-going channels of communication with them**. This is likely to work best where the meta-evaluation is seen to be adding value other studies, as well as vice versa.

The implication of the Meta-Evaluation team's experience is that **it is essential that meta-evaluators undertake systematic assessment of the quality and relevance of secondary evidence**. These assessments can be used in two ways**. Where there is plenty of data and evaluation evidence, quality assurance identifies the most reliable sources**. **When there is only one source of evidence, quality assurance should be used by meta-evaluators to identify those conclusions which rely on less reliable evidence and should therefore be given a 'health warning'**.

Another clear pointer from the Meta-Evaluation of the 2012 Games is that in the real world meta-evaluation will often be as much about **synthesis of research findings** as it is about aggregation of data from other studies and surveys, which may not be feasible. Synthesis requires the meta-evaluator to weigh up the findings/conclusions of other studies in order to reach a **judgement** about 'high level impacts'. This is an area in which there would be benefit in **work to develop meta-evaluation skills and tools.**

Our review of other meta-evaluations commissioned by the UK government and studies of other mega-events supports this view. Most of the studies we identified had experienced similar challenges and it seems clear that **the ability to aggregate data from diverse sources will be rare**, even in meta-evaluation. **It is made very difficult if there is no effective planning and orchestration of studies. Meta-evaluators could play a role in enabling this and this provides another reason why they need to be in place early in the overall evaluation process.**

Where it is not feasible to commission meta-evaluations ahead of other studies, it would be **advisable to allow a contingency budget to allow for additional work to fill gaps and analyse unanticipated policy developments and outcomes**. This was made easier in the case of the Meta-Evaluation of the 2012 Games because of interest in the Olympics – other meta-evaluations will not necessarily find it so easy to initiate (or influence) primary research, where there are gaps in the data or some of their expected component evaluations fall by the wayside. It was also suggested that it would be helpful to **make greater use of other types of evidence, for example expert judgement might be deployed to complement 'hard' data**. Moreover, performance management information from public programmes might be useable in more imaginative ways to help inform both formative and summative evaluations; meta-evaluation methods could be valuable both to assess the quality of such data and how best to integrate them into the overall assessments.

Counterfactuals

The scale and complexity of mega-events make constructing the counterfactual position particularly difficult. This is even truer in the case of one-off events like the 2012 Games. Problems producing robust counterfactuals are not, of course, unique to meta-evaluation. But it can be doubly hard for meta-evaluations, since they are **reliant to a large extent upon counterfactual methodologies being deployed in other evaluations**. These are likely to differ in their scope and level of robustness, to the extent that they are used at all. Interviewees nonetheless considered that **if meta-evaluation is worth doing, then an attempt at constructing counterfactuals is usually going to be worth trying, even though it will, by definition, be difficult and less than perfect.** This may involve reanalysing secondary evidence in different ways, conducting further new primary research, and building additional variables into nationally representative surveys, all of which were attempted for the Meta-Evaluation of the 2012 Games. This reinforces the desirability of **starting meta-evaluations early in order to give them the best possible chance of capturing baselines and monitoring relevant trends, and helping to set the parameters for counterfactual assessments**.

Longer-term impacts

Meta-evaluations tend to deal with complex issues, whose full consequences only become evident over a long time period. This was clearly true in the case of the legacy of the 2012 Games – here the difficulties inherent in getting commitment to future follow-up research into longer-term impacts were recognised from the outset, and some steps were taken to deal with this issue. For example, legacy questions have also been introduced into national surveys which will be repeated regularly in the future and, in addition, there is a commitment by government to provide an annual update on the legacy of the Games to Parliament, which will require compiling latest data from relevant surveys and evaluations completed since the final Meta-Evaluation report. This illustrates how meta-evaluations should design their own legacy in the shape of sustainable approaches to the exploration of longer-term impacts.

Independence of the meta-evaluation

Both policy makers and the research team have emphasized that the greatest benefit of the Meta-Evaluation has been its 'quality mark' as a rigorous, independent evaluation. This enabled it to provide an important mechanism for accountability. Having a cross-government Steering Group helped, by providing 'checks and balances' between government departments. However, it may be that **further mechanisms to ensure independence should be built in to meta-evaluations in future** (e.g. having an independent body to oversee all major government evaluations or to provide the chair of the Steering Group). *This would follow the precedent set by the development of the UK's Office for Budget Responsibility, and the development of the new What Works Centres established jointly by ESRC and the UK government.*

Using the meta-evaluation
Policy makers acknowledged that they tend to have only a relatively thin interface with research, and quite mixed motives for using research evidence. Often they use it to pluck out favourable findings to promote their own policies and, inevitably, one of their main interests in a meta-evaluation is to ensure that their own activities have been positively reviewed and reported. The 2012 Games Meta-Evaluation team identified a wide range of potential users of and uses for their work, and interviewees were optimistic that its findings would be taken up.

Of course, the Meta-Evaluation of the 2012 Games was unique in many respects – and all meta-evaluations are similarly likely to have unique elements. However, the perceived value of the 2012 Games Meta-Evaluation flags up the potential for meta-evaluations in other government programmes – something which has been tried relatively rarely in the UK government. Its added value came from adding an extra, broader level of analysis and interpretation to the 'component evaluations' on specific aspects of the Games – as well as, albeit to a more limited extent, influencing the component evaluations in such a way as to make their findings more complementary. These benefits seem relevant to a wide range of government programmes, where traditionally project-based evaluations have been commissioned but the bigger picture has not been brought together. **Both policy makers and the research team strongly recommended that more opportunities be sought for conducting meta-evaluations into UK government policies.**

Skills and training
If meta-evaluations are to become more common in the future, this suggests that **research councils and/or others should seek to enhance the capacity for meta-evaluation within the UK research community. In particular, they should pay attention not just to training in methods but also to developing research leadership and stakeholder management skills**.

Summary conclusions

1. The research literature and interviews with a range of experts revealed that there is a variety of methods and terminology and lack of clear methodology to describe the evaluation of mega events. Meta-evaluation is one way to describe such evaluations.
2. There are three main forms of meta-evaluation: (i) meta-theory on the meaning of evaluation; (ii) the assessment of the quality of evaluations; and (iii) a form of synthesis of individual evaluations.
3. Meta-evaluation as synthesis best fits the evaluation of mega-events and is a special form of impact evaluation systematic review where sub-components (rather than the more commonly found variable instances) of an intervention are brought together to answer a research question.
4. Impact evaluation systematic review methods thus provide a basis for a methodology for the impact evaluation of mega-events.
5. As with all research and research synthesis, the nature of the research question and theory are both crucial in determining fit for purpose research methods. Particular challenges of such methods for mega-events such as the Olympics include:
   - The number of sub-questions including cross cutting questions that need to be answered to answer the macro impact questions

- The difficulty in creating or finding the data necessary and the likely heterogeneity and variable quality and relevance of that data
- The importance of theories or logic models to structure the evaluation and to manage and analyse the diverse range of data to answer the macro and sub-questions
- The need for possible iteration in both the theoretical framework and research data as the evaluation progresses
- The challenges of trying to identify causal attribution when there is unlikely to be much experimental control of large data set models to test hypotheses, or other counterfactual methodologies within component studies.
- The importance of time in: (i) planning the evaluation including the research that provides necessary data, prior to the mega-event (both baseline and process and outcome data); (ii) the ability to assess long term legacy outcomes and legacy impacts.

6. To enable a high quality meta-evaluation and to make its policy impact more likely, there is a need for a policy customer who is interested in the overall picture and can influence other departments and agencies, and also for a clear guarantor of independence for the meta-evaluation, perhaps in the form of an independent chair of the steering group.
7. These lessons of meta-evaluation from the 2012 Games are relevant not only for other mega-events but also for evaluating broad multi-component policy initiatives. Such meta-evaluations would benefit from developments in the policies and infrastructure for research in support of policy making (policies in research production and use) as well as in meta-evaluation capacity.

1: Introduction

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

1.1 Background to the study

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 London 2012 Olympic Games and Paralympic Games. The study was of the utmost importance in demonstrating the legacy and impacts of the 2012 Games across all thematic areas to the end of 2012, and was the single largest and most comprehensive evaluation exercise commissioned in connection with the event up to that point. The study was to 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 ..." and also "…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 London 2012 meta-evaluation therefore held 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 London 2012 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

[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.

and meta-synthesis research, to develop a research specification and assist with conducting the research. The research team included at various times:

- 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, Director of the Public Policy Institute for Wales and 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.
- Tony Bovaird, Professor of Public Management and Policy, Institute of Local Government Studies; and Third Sector Research Centre, University of Birmingham.
- 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 the London 2012 meta-evaluation project leads at Grant Thornton and Ecorys 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 18 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 and configuration 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. Section 3 provides 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 Olympics or any mega-event can thus be assessed, prior to the synthesis of empirical results and aggregation of the overall impacts on beneficiary groups and 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 London 2012 meta-evaluation, 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 Impacts and Legacy of the London 2012 Olympic Games and Paralympic Games, which can advance methods of meta-evaluation?

The methodology 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 London 2012 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 Games. This review is included in sections 2 and 3 of this report.

Roundtable discussion on methods: two roundtable discussions were convened between the academics, operational members of the meta-evaluation team and DCMS. 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 London 2012 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 was undertaken 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 4 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 in the London 2012 meta-evaluation, in relation to the collation, review and synthesis of sources of evidence and the reporting of results (section 5).

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 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 in the review 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. The studies are therefore syntheses (the third type of meta-evaluation identified in the following chapter). However, they do highlight some important methodological issues which are relevant to the London 2012 meta-evaluation.

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:

- most mega-events have multiple objectives;
- their stated objectives evolve over time;
- different groups articulate different kinds of objectives; and
- outcomes may be negative and/or 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:

1. 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 through sports participation.
2. 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.
3. 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 needed to be taken into account in the London 2012 meta-evaluation. 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 focus 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); and
- Boosting participation in sport and enhancing sports infrastructure (championed by both the previous Labour administration and the current Coalition Government, sports bodies such as Sport England, and sportsmen and women themselves).

The aspirations attached to different mega-events also reflect the wider political and economic 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 seen as important because they took place as Spain emerged from a period of dictatorship and also entered the EU. Similarly, the Rugby World Cup was regarded as a defining moment in post-apartheid South Africa.

In recognition of the often multiple objectives and scale of mega-events, 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 evaluations recognise other types of impact or legacy as important, though they rarely agree on what these are. Legacy themes or indicators used in previous studies include:

- Improvements in governance capacity;
- Promotion of national and/ or regional identities;
- 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 theme or indicator 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:

1. Conception;
2. Organisation;
3. Staging; and
4. Closure.

The Rand Corporation (undated) suggests using three periods:

1. Planning;
2. Delivery; and
3. Legacy.

It may be that different kinds of impact measures and/or (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 and researchers to make sure the data they need will be gathered, and conducting a formative assessment of impact.
- During the implementation phase evaluators may be engaged in data gathering to help assess the short-term and immediate impacts of staging the event, whilst working with other evaluators and researchers 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.

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 cannot 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').

Clearly there is a difficult trade-off to be made. To take too broad and too long a view on possible legacies and impacts would risk undermining the reliability and credibility of any meta-evaluation. But to focus too narrowly would be to miss many of the anticipated benefits of the Games which are by nature indirect and possibly even intangible (Langen and Garcia 2009).

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).

The unintended impacts and consequences of mega-events are frequently also a focus of studies. It may be legitimate for evaluations to explore the extent to which such potential impacts are anticipated, planned for and reacted to when they occur. Some studies emphasise the importance of including locals' views in evaluations of mega-events, and some experiment 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 the events.

2.6 Integrating evaluative frameworks

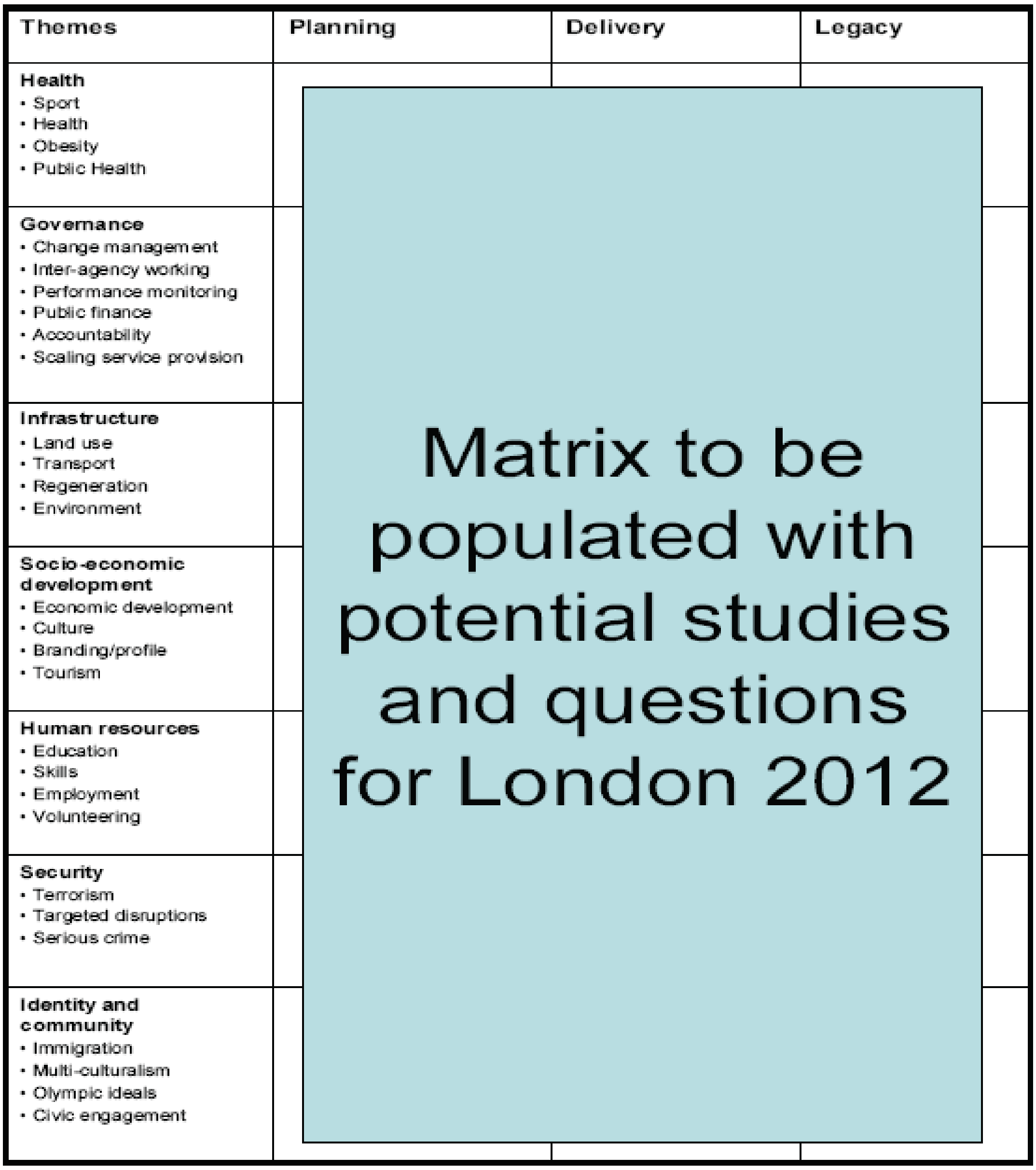
Different kinds of mega-event impacts and legacies require different measures and possibly evaluation methodologies, so it is challenging to find a grand conceptual amalgam capable of reflecting all ambitions.

The literature nonetheless offers some possible pointers to frameworks that might help to structure the Meta-Evaluation of the Impacts and Legacy of the London 2012 Olympic Games 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 also 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 we discuss in the following chapter.

Figure 1: Evaluation matrix for mega-events



Themes	Planning	Delivery	Legacy
Health • Sport • Health • Obesity • Public Health			
Governance • Change management • Inter-agency working • Performance monitoring • Public finance • Accountability • Scaling service provision			
Infrastructure • Land use • Transport • Regeneration • Environment			
Socio-economic development • Economic development • Culture • Branding/profile • Tourism			
Human resources • Education • Skills • Employment • Volunteering			
Security • Terrorism • Targeted disruptions • Serious crime			
Identity and community • Immigration • Multi-culturalism • Olympic ideals • Civic engagement			

Matrix to be populated with potential studies and questions for London 2012

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.

3.1.1 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.

3.1.2 Meta-evaluation of the quality 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 assessment 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)

3.1.3 Meta-evaluation as synthesis of findings

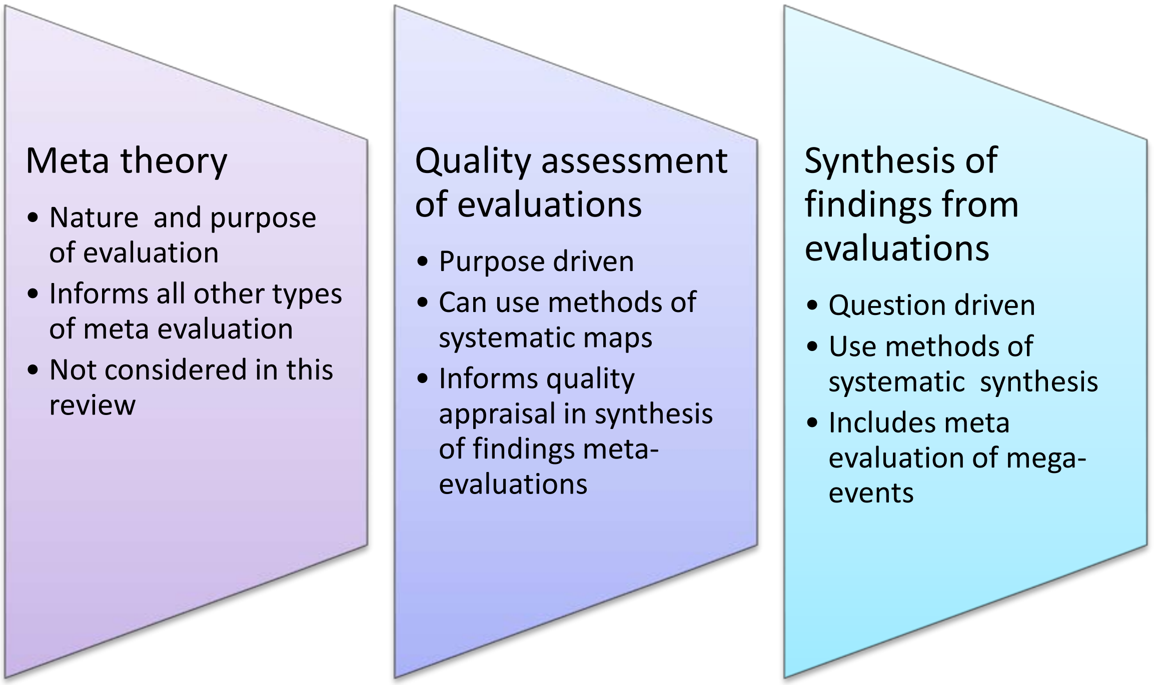
Another type of meta-evaluation is the synthesis of the findings from individual studies to answer an evaluation research question. In other words, this is the combination (or aggregation) of multiple evaluation studies. Evaluation is often of an individual occurrence of a single intervention. In meta-evaluation, 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 London 2012 meta-evaluation, 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 assessment of evaluations and synthesis of findings from evaluations.

Figure 2: Three Main Types of Meta-Evaluation



Meta theory

- Nature and purpose of evaluation
- Informs all other types of meta evaluation
- Not considered in this review

Quality assessment of evaluations

- Purpose driven
- Can use methods of systematic maps
- Informs quality appraisal in synthesis of findings meta-evaluations

Synthesis of findings from evaluations

- Question driven
- Use methods of systematic synthesis
- Includes meta evaluation of mega-events

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 Impacts and Legacy of the London 2012 Olympic Games 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 the quality assessment of evaluations and the synthesis of findings from 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 assessment 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).

3.3.1 Aims and methods of quality assessment 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.

3.3.1.1 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.

3.3.1.2 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 boxed example 1 from Cooksy and Caracelli). Organizations 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, the Department for International Development in 2008 reviewed the evaluation methodology used in its 'country studies' and sought to strengthen the methodology, using experience from comparable evaluations in other parts of the UK Government and internationally).

Example 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 example 2 on Oliver et al.. 2010).

Example 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 example 3).

[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.

Example 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)

3.3.1.3 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 example 4 from Yarbrough 2011) and the Evaluation Centre at Western Michigan University[4] plus many others in the United Kingdom[5] and further internationally.

Example 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.

[3] http://www.jcsee.org/

[4] http://www.wmich.edu/evalctr/checklists/

[5] http://www.evaluation.org.uk/resources/guidelines.aspx

3.3.2 Dimensions of difference in quality assessment 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 18 meta-evaluations of single studies Cooksy and Caracelli (2009) found that five were assessed according to quality standards, three using criteria developed specifically for that meta-evaluation, three used the criterion of trustworthiness based on the confirmability and dependability of the findings, and seven 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

Although there are many types of quality assessment meta-evaluations, 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 example 5).

Example 5: 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)

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

- **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).

Whilst being mindful of such differences, Stufflebeam suggests the following broad steps for carrying out quality assessment meta-evaluations (2001, pg. 191):

Figure 3: Flexible Structure for Undertaking Quality Assessment Meta-evaluations

1. Determine and arrange to interact with the meta-evaluation's stakeholders
2. Staff the meta-evaluation team with one or more qualified evaluators
3. Define the meta-evaluation questions
4. Agree on standards, principles, and/or criteria to judge the evaluation system or evaluation

5. Develop the memorandum of agreement or contract to govern the meta-evaluation
6. Collect and review pertinent available information
7. Collect new information as needed, including, for example, through on-site interviews, observations, and surveys
8. Analyse the qualitative and quantitative information
9. Judge the evaluation's adherence to appropriate standards, principles, and/or criteria
10. Convey the meta-evaluation findings through reports, correspondence, oral presentations, etc
11. As needed and feasible, help the client and other stakeholders to interpret and apply findings

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 all existing research studies relevant to 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, selection and synthesis of study findings to answer the review question (Gough and Thomas 2012). This objective and transparent approach is adopted in order to minimize bias in the review findings. Systematic reviews can be both quantitative and qualitative in nature, or a mixture of both, and can have various aims and employ various methods of synthesis.

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

3.4.1.1 Aggregating and configuring reviews

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 impact and 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.

3.4.1.2 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 example 6 from 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.

Example 6: 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)

3.4.1.3 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 (both quantitative and qualitative). 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 example 7 from Pawson 2002).

Example 7: 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 that is 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)

3.4.1.4 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 example 8 from Britten et al.. 2002).

Example 8: 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 example 9). 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.

Example 9: Diffusion of innovations in health service organizations

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)

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, as illustrated below. Here qualitative synthesis is employed alongside quantitative synthesis to help explain heterogeneity in impact results, and to develop theories of why interventions are successful and how they can be improved, based upon the identification of confirmatory/contradictory patterns in the evidence.

Example 10: Barriers and facilitators 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: Using conventional systematic review methods, 33 experimental trials and eight qualitative studies of children's views were found that met pre-specified inclusion criteria. The studies were assessed for quality and reliability according to standards for their specific type of study; they were then synthesised individually using methods appropriate to the study. The studies were assessed in terms of reporting quality, internal validity or reliability, and for qualitative studies the extent to which the findings were rooted in children's perspectives. Nineteen outcome evaluations were considered to be sufficiently reliable to enter into a statistical meta-analysis, in order to synthesize the quantitative data from these studies and estimate the scale of impact (as well as identifying heterogeneity). The textual findings from eight qualitative studies were then analysed through a thematic synthesis, to identify barriers and facilitators based upon children's perspectives and understandings. The findings of both syntheses were then brought together to answer the main review question, through assessing whether interventions which matched children's views were more effective than those that did not. This resulted in one review with three syntheses.

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 that were in line with these suggestions tended to be more effective than those that were not.

(Thomas et al. 2004)

3.4.1.6 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 example 11 from Caird et al.. 2010).

Example 11: 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)

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 example 12 from 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.

Example 12: 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. This might include for example 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.

Example 13: Meta study on decentralized 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 or 'joined-up policy-making' is 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.

3.4.2 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 (for more details see Gough and Thomas 2012; Gough et al.. 2012). This section identifies some of the main dimensions of difference in synthesis meta-evaluations.

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 likely to be 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.

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.
- **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.

The Cochrane Collaboration provides a handbook for systematic reviewers of interventions (Higgins JPT, Green S (editors), updated March 2011). This outlines eight general steps for preparing a systematic review (Figure 4).

Figure 4: Flexible Structure for Undertaking Synthesis Meta-evaluations

1. Defining the review question and developing criteria for including studies
2. Searching for studies
3. Selecting studies and collecting data
4. Assessing risk of bias in included studies
5. Analysing data and undertaking meta-analyses
6. Addressing reporting biases
7. Presenting results and "summary of findings" tables
8. Interpreting results and drawing conclusions

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 from evaluations. It is the third of these, **synthesis of findings from evaluations**, which best describes the impact meta-evaluation of mega-events such as the Olympic and Paralympic Games and by extension other major public policy interventions.

The synthesis of findings from evaluations 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), to deepen knowledge of issues such as impacts and process lessons. In the impact meta-evaluation of a mega-event, however, it is relevant evaluation studies from different (but nonetheless related) subcomponents of the same event that need to be brought together and synthesized, in order to provide a fuller understanding of the outcomes of the event and its longer-term legacy. This implies additional layers of complexity; simultaneous sub-component meta-evaluations at various levels must first be carried out before the results of these thematic syntheses are themselves combined to answer overarching questions about the impact and lessons from mega events.

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 detail of the methods used. The result is that the literature is rich on conceptual issues, though no paper is comprehensive, but thin on technical issues.

The methods used to (meta) evaluate a mega-event or other intervention can, however, follow some of the methods of **systematic review**; this can help to minimise bias and ensure transparency of findings. There is a very rich detailed literature on systematic reviews, which are relevant to 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. Section 3.4 nonetheless sets out the different types of systematic review and methodologies in broad terms, which could be considered relevant to synthesis meta-evaluation.

If an overarching categorisation is sought, then the impact meta-evaluation of a mega-event might be considered to be closest, in its aims and approach, to a '**mixed-methods, aggregative theory testing review**'. This would involve testing the efficacy of groups of interventions and phenomena relating to a mega-event's impact and legacy against a pre-constructed (but malleable) set of programme theories and concepts, assumptions and measures, based upon data collected in a relatively systematic way.

Many of the papers on meta-evaluation 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 their accompanying guidance are of great value to meta-evaluations of public policy and in particular 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. As part of systematic review, the quality assessment of component studies is an integral element of synthesis and reporting; it can help weight and strengthen the claims made by the study.

In the case of the London 2012 meta-evaluation, which incorporates both formative and summative stages, multiple purposes for quality 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);

- 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.

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 review and **synthesis methods** that need to be combined to address one or more overarching questions (St Pierre 1982). In general terms though, syntheses of impact or other quantitative findings are brought together with syntheses of qualitative research with beneficiaries and other stakeholders, to help test the theories of change and answer the mega-event meta-evaluation questions, so far as confirmatory or contradictory patterns are available in the evidence.

Finally, some 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 conducting a successful meta-evaluation of a mega-event.

Section 5 builds on the findings from the literature review, and relevant methods of systematic review in particular, to provide a specific set of guidelines for structuring and implementing the synthesis methodology of the Meta-Evaluation of the Impacts and Legacy of the London 2012 Olympic Games and Paralympic Games. This will also be relevant to other impact meta-evaluations of mega-events and of any complex government policy or programmes. First of all in section 4, we consider the perspectives of recent practitioners of meta-evaluation, to help incorporate transferable lessons and evidence of what works from the research community.

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) 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 Appendix 3). 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

4.3.1 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 type 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.

4.3.2 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.

4.3.3 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.

4.3.4 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 the impact and use made of evaluations. They advised that this is the commonly understood definition of meta-evaluation in the European evaluation community. One described it as the:

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

The other saw meta-evaluation as:

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

Methods for this kind of study were, they said, well understood and were presented to European evaluation standards.

4.3.5 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-evaluation 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)

Meta-evaluation can also help understand 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.

4.4.1 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.

4.4.2 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 achieving significant reductions in 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 a 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 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).

4.4.3 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)

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

4.4.4 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.

4.4.5 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 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)

4.4.6 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’, (complex societal problems such as unemployment, ill health and environmental damage) 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 London 2012 meta-evaluation

Turning to the Meta-Evaluation of the Impacts and Legacy of the London 2012 Olympic Games 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 host 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 of legacy and there are examples in the literature of evaluations which include these impacts, but others argued that these were too ill defined in general 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 London 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 cautioned that the stated objectives of the London 2012 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 London 2012 meta-evaluation and other evaluations of mega-events.

The results of the interviews 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 any meta-evaluation**. This places the London 2012 meta-evaluation firmly in the synthesis camp. It has drawn on evidence from a range of sources including other evaluations and needed to test whether these secondary data sources were reliable. However, the primary task was 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 also **define the nature of the legacy which mega-events such as the 2012 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 London 2012 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 as being clear about the type (or types) of legacy, **it is** important to **be clear about the distribution of legacy**. Any meta-evaluation of a mega-event should attempt to assess which areas and which sections of society benefit (or experience dis-benefits).

Fourth, it is important to know not just whether but also **how legacy is achieved**. Several of the experts recommended developing a theory-based approach which models the ways in which Olympic and Paralympic 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 meta-evaluations of mega-events should **focus on 'high level outcomes', and pay attention to interactions** between different aspects of the events such as the Olympic and Paralympic 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 London 2012 Games, because the data were unlikely to be available. They advised that the London 2012 meta-evaluation would need to take a pragmatic approach, **drawing upon a range of very different kinds of evidence**, but also looking 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, they suggested that it would be difficult to identify clear baselines or counterfactuals for the 2012 Games. Establishing cause and effect mechanisms would not therefore be straightforward, and time lags could mean that the full extent of any legacy was 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 meta-evaluation efforts 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

As previously stated, the impact meta-evaluation of mega-events most closely resembles the synthesis of evaluations form of meta-evaluation. Indeed, an early step for the London 2012 meta-evaluation was to determine how relevant evaluations and their results were to be identified and integrated, in response to the overarching research objectives. The wide variety of impacts identified for any mega-event 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 of a mega-event.

In the previous sections it was concluded that this requirement shares many of the characteristics of a multi-component, mixed-methods systematic review. The synthesis includes empirical outcome data. It also involves 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. 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.

This chapter builds upon this understanding to provide guidelines for the generic stages of a mega events impact meta-evaluation (rather than a description of the particular steps and strategies undertaken for the London 2012 meta-evaluation, although this guidance informed the approach).

The London 2012 meta-evaluation was driven by a set of logic models and theories of change hypothesizing how the Games might impact on four broad types of outcome. 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 5 (the process may be iterative but is shown as a linear list here for clarity). Although this was designed to help structure the methodology for the Meta-Evaluation of the Impacts and Legacy of the London 2012 Olympic Games and Paralympic Games, it also has more universal applicability.

Figure 5: 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 requireda**
   - **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**

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**

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

These stages and steps are explained in more detail below. A number of example research tools, drawn from the London 2012 meta-evaluation, are referenced in the text and can be found in appendix 5.

5.2 Stages of an impact meta-evaluation

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 (Step 1.1 and 1.2). 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, the 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 and mechanisms by which positive or negative outcomes are thought to occur (sometimes called logic modelling), and the underlying assumptions involved (sometimes called theories of change), prior to the specification of relationships between overall questions and models and the various sub-questions and their models (Step 1.3).

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.

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 (and associated indicators of success: see Stage 2). 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 also likely to have differential effects if provided in different ways to different groups in different situations; the theories can 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).

The research strategy for these impact meta-evaluations can therefore be no better than the concept maps which precede them. 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. Nonetheless, the theories and their logic models 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.

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), the 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 theoretical outcomes rather than 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 negative effects) may not be captured within each theme. Logic models nonetheless also lend themselves to this process since they provide the basis for linking data together beyond individual interventions and their outcomes.

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. Cross cutting issues can thus be questions that are asked in advance, or which arise iteratively as the data is examined and the empirical data is considered against the prior hypotheses and detailed logic models.

These thematic maps then form the basis of an evaluative framework (Step 1.4) for considering meta-evaluation questions, which in turn inform the specific methods of meta-evaluative review that need to be applied to identify, appraise and synthesize the relevant evidence. 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. This process can be described as follows:

1. **Selection of overall meta-evaluation questions**
   (i) Process of selecting stakeholders and involving them in question selection
   (ii) Deciding on overall questions, cross-cutting questions and consideration of other questions not selected
   (iii) Identify perspectives and assumptions (theoretical and ideological framework) including any model within the questions being considered
   (iv) The review methods used to identify, appraise and synthesize the relevant evidence

2. **Selection of sub-questions**
   (i) Process of selecting stakeholders and involving in sub-question selection
   (ii) Deciding on sub-questions and consideration of other questions not selected, including how they answer the overall question and cross-cutting questions;
   (iii) The perspective and assumptions (theoretical and ideological framework) including any model within the sub-questions and how they relate to each other and to the overall questions and framework including ‘process’, ‘outputs’ and ‘outcomes’; these can include cross cutting questions that cut across the major questions and sub-questions, some of which may arise iteratively as the meta-evaluation progresses.
   (iv) The review methods used to identify, appraise and synthesize the relevant evidence.

These theories, concepts and questions in turn inform the specification/types of evidence sought (Stage 2). Inappropriate problem formulation at this stage is a major risk. If the research questions are not clear then it is unlikely that they will be operationalized in the research study in a way that the study will be able to answer them. 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 were therefore critical elements of the scoping phase of the Meta-Evaluation of the Olympic and Paralympic Games.

Appendix 5 includes an example logic model and question set from the community engagement theme of the London 2012 meta-evaluation.

STAGE 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 (Step 2.1).

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 (Step 2.2). The particular methods used to search for evidence that fits these criteria will similarly be framed by the type of review being applied. For example, this determines 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) which can facilitate 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). Relevant data sources can then be identified (Step 2.3) 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 need to be initially checked (screened, Step 2.4) and then compared against the data required (inclusion criteria, Step 2.5), before consideration is given to whether further data should be sought (Step 2.6). This may include the commissioning of further primary research. The data that is available is, of course, limited by the studies available. In the case of the Olympics, this may include primary evaluation studies set up specifically to evaluate Games components, other studies that have happened to have been undertaken and are relevant, and on-going and one-off surveys to inform the synthesis. The studies may provide data on change subsequent to a mega-event, and/or data to provide evidence of additionality (to control for counterfactuals). There may also be many gaps.

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:

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

2. **For the overall or top level questions:**
   - (i) To what extent will the sub-questions provide answers to the overall questions?
   - (ii) Identifying what further data are necessary

Example evidence tables, drawn from the volunteering and social action sub-theme of the community engagement theme, are provided in appendix 5.

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 (Step 3.1). 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 (Step 3.2). A third reason is to be able to describe (or 'map') the nature of the research field of research evidence meeting the inclusion criteria (Step 3.3.); 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 (Step 3.4) 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 (Step 3.5).

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).

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 London 2012 meta-evaluation, for example, had not been of good quality or relevant then the findings and conclusions of the Meta-Evaluation may not have been valid.

The syntheses are 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 synthesis 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) also 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 aim of promoting quality in evaluation practice (and hence incorporates significantly more detail than was required by the London 2012 meta-evaluation) some of the specific standard statements are relevant to the meta-evaluation of mega events.

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 evaluation of studies for inclusion in the synthesis therefore depends on three main dimensions (Gough 2007) including not only the technical quality of the evaluation (Step 4.1), but also the fitness for purpose of that method for the review (Step 4.2), and the validity of the approach used in the study relative to the review question (4.3). 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

[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

dimensions B and C (WoE B and C) are (meta) evaluation specific criteria. The weight of evidence system allows the meta-evaluator to assess the worth of the study in answering the review question(s). This is not necessarily the same as assessing a study on its own as the aim of the study may not exactly fit the aims and perspectives of the meta-evaluator.

In practice, these dimensions are applied in different ways. In terms of the first dimension of technical execution (Step 4.1), 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:

1) Are the aims and objectives of the research clearly stated?
2) Is the research design clearly specified and appropriate for the aims and objectives of the research?
3) Do the researchers provide a clear account of the process by which their findings were reproduced?
4) Do the researchers display enough data to support their interpretations and conclusions?
5) 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 (Step 4.2), 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 (Step 4.3), this is also often determined by inclusion criteria and by later reviewer judgements of adequacy. An example, relevant to sports-related 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 (Step 4.4). 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

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

use outcome measures that are not very valid for the synthesis. An example of this process is given in Appendix 2.

Given the wide ranging purposes of quality appraisal associated with the London 2012 meta-evaluation (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 need to be combined to produce a suitable appraisal tool for mega-events meta-evaluation.

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 potentially 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 London 2012 meta-evaluation, the individual studies were mostly undertaken by other researchers. Where members of the Meta-Evaluation team were involved in the generation of data for synthesis, this was to complete a missing part of the knowledge needed for the Meta-Evaluation, or else there was a strong separation from the evidence appraiser.

A summarised version of the Quality Assessment (QA) tool employed by the London 2012 meta-evaluation team is included in appendix 5.

Stage 5: SYNTHESIS

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

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

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

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 (either confirmatory or contradictory) in support of all of 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 al.one the combination of several sub-questions to address headline questions. Synthesis has a number of practical stages.

First is clarification of the data available to be interrogated to answer the review questions (Step 5.1). 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 Games and Paralympic Games, this process is likely to yield a wide and disparate range of information for synthesis including: outcome and process evidence from focused evaluations of specific interventions; 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.

The review question is then used to drive the examination of patterns in the data (Step 5.2). 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.

There may also be benefits in searching for new patterns that might exist in the data (configuring rather than aggregating data). This is a post hoc rather than a priori testing of hypotheses. This may also include configuring patterns in concepts or theories about the phenomena being studied.

When the data is very varied, the process of seeking patterns may require mixed methods of data aggregation and conceptual configuring 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.

Synthesis may also involve sub-component syntheses where different aspects of an issue have been interrogated by sub-questions (Step 5.3). 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.

Step 5.4 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.

Stage 6: CONCLUSIONS AND DISSEMINATION

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

Step 6.2: Interpret and test findings

Step 6.3: Assess strengths of the review

Step 6.4: Assess limitations of the review

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

Step 6.6: Refine theories in light of evidence

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 (Step 6.1). 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 (Step 6.2). 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 (Steps 6.3 and 6.4). 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 each major question (Step 6.5). 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. For the London 2012 meta-evaluation there were issues of quality appraisal for each of the stages of the work.

Once the results of a meta-evaluation have been interpreted, tested and qualified (including through the process of refining initial theories, Step 6.6) 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 (Step 6.7).

6: Learning from the Evaluation Team

6.1 Introduction

The Meta-evaluation of the Impacts and Legacy of the London 2012 Games was an ambitious undertaking which provided a valuable opportunity to learn more about how meta-evaluation operates in practice. The way in which the study was designed, the challenges it posed, the methodologies developed by the Meta-evaluation team and the issues met in implementing these methodologies provide important lessons for future research. The project also represented an opportunity to reflect on the strengths and weaknesses of meta-evaluation as a means of analysing the impacts and legacy of mega events.

We gathered evidence from the lead members of the team that undertook the study by convening workshops at key points in the study and a series of semi-structured interviews at the end of the project. The workshops served two purposes. They gave the Meta-evaluation team access to academic advice about the methodological issues they were grappling with, including in particular the:

- Development of theories of change and logic models;
- Quality assurance of secondary evidence; and
- Synthesis of evidence from other evaluations and surveys.

They also enabled us to study, in real time, the methods the Meta-evaluation team was using and the challenges which they encountered.

The interviews were conducted at the end of the study between May and July 2013. They enabled us to explore the team's reflections on its work and the lessons for future meta-evaluations. We interviewed the team leader, lead evaluators for each of the main themes, and the project manager who oversaw the Meta-evaluation for the Department for Culture, Media and Sport. All of the interviewees had been closely involved in the Meta-evaluation and provided detailed information about the challenges involved in the study and what they saw as the advantages and disadvantages of the methods they had employed. Interviews were structured using a topic guide (included in appendix 4) which was informed by the main themes identified by our reviews of the literature on meta-evaluation and mega events (chapters 2 and 3). They were conducted one-to-one, in person, on a non-attributable basis and lasted an average of around 60 minutes. They were recorded using tape and/or contemporaneous notes and analysed using a matrix based on the topic guide structure.

6.2 Objectives of meta-evaluation

Our literature review (chapter 3) showed that there are very different schools of thought within the academic community about what meta-evaluation involves. Our interviews with leading evaluation practitioners (chapter 4) found a similarly diverse range of interpretations of and approaches to meta-evaluation, ranging from standard setting to impact assessment to meta-analysis and synthesis studies.

Some of the Meta-evaluation team were unaware of the term 'meta-evaluation' prior to the study, but they all had the same understanding of the project. They saw their job as being to make an overall assessment of the impacts and legacy of the London 2012 Games by bringing together diverse evidence drawn primarily from secondary sources. There was some provision in the Meta-evaluation budget for primary research, but both the DCMS and the Meta-evaluation team expected that the bulk of the evidence would come from other evaluations and surveys focused on specific types of legacy and impacts. As one interview explained, the Meta-evaluation was designed as:

> An 'evaluation of evaluations' (which would) draw together the evaluations ….. and bring together different pieces of evidence.

Members of the team thought that the Government had been right to seek this overall assessment of the impacts and legacy of the Games and believed that the objectives set by the DCMS for the study were appropriate. They also agreed with the research questions that they had been set, and believed that the Department had been right to emphasise from the outset that the Meta-evaluation would provide an interim assessment, rather than the final word on impact and legacy.

An important lesson for future research is that in spite of the considerable challenges involved in undertaking it (which we explore in detail below), **meta-evaluation is seen as having an important role to play in enabling an overall assessment of the impacts and legacy of high profile, large scale interventions and events** of the kind that can not be provided by more narrowly focused evaluations of individual projects and programmes.

6.3 Thematic approaches and cross cutting issues

One of the challenges of the Meta-evaluation was the scale, scope and complexity of the study. There were a very large number of potential research questions and multiple interconnections between different kinds of impacts and legacies.

The Meta-evaluation addressed this by structuring the study around six key themes (which were later reduced to four[12]). The Meta-evaluation team members believed these were the right themes and a good way of organising the study. They had made the project manageable and provided clarity about the focus of the study. It also helped with the organisation and management of the team since members could be allocated clear roles and responsibilities and were able to specialise in particular kinds of impacts and legacies. Indeed, interviewees found it hard to think of an alternative way of organising the team and most said they would advocate a similar approach to future studies. They told us:

> We needed the thematic approach to make the task manageable.
>
> It worked well as an organising approach and the themes stood the test of time.

However, they also pointed to some drawbacks of this approach. They regretted the reduction from six to four themes part way through the study. They felt that the thematic approach risked locking them into an approach which made it difficult to respond to changes in government priorities. And they found that it had been difficult to take full account of linkages between themes. They had sought to analyse a small number of 'cross-cutting' issues that were relevant to all of the themes including the legacies in terms of disabilities and sustainability, and this approach was seen as having worked reasonably well. However, the team had not been able to pull together evidence about cross cutting issues until the latter stages of the study. As one interviewee reported:

[12] Sport, Economic, Community and East London, all incorporating aspects of sustainability and disability which started out as separate themes

> It worked fine but the cross-cutting issues definitely got less attention. The (core) themes definitely got more attention, particularly in terms of primary research.

As a result there was a risk that cross-cutting issues:

> became a bit of an afterthought… (we) could have invested more time in the management of this part of the study.

One of the lessons from the Meta-evaluation is, therefore, that it is **important to focus on core themes which provide a clear focus for a meta-evaluation but to combine this with an awareness of issues which cut across themes.** The Meta-evaluation team believed that future studies would do well to adopt a similar approach to the one used for the London 2012 Games. But some interviewees recommended **designating a senior team member to take specific responsibility for championing 'cross-cutting' issues and ensuring they received sufficient attention throughout the meta-evaluation**.

6.4 Logic models and research questions

In addition to the themes and cross-cutting issues, the team used logic models to identify the specific research questions that they would address within each theme. Interviewees reported that this approach proved very effective. It was:

> Very helpful in framing the way we addressed the research questions.

The logic models:

> gave a thread to follow through.

and acted as:

> an organising principle.

This helped the team to identify the key issues and to stay focused on them.

However, they identified three risks associated with the use of logic models. First, they were designed before the team knew what data were going to be available and this meant that in practice it was difficult to follow through some of the hypotheses about cause and effect. Second, the objectives (and therefore expected outcomes) of the Games shifted over time, partly because the Coalition Government had some different priorities to the Labour administration that had been in office at the beginning of the Meta-evaluation. This meant that some issues were not reflected as fully as they might have been in the original logic models. One interviewee explained, for example, that:

> the emphasis on volunteering came after the bid. It was grafted together with emerging ideas on Big Society.

Another told us:

> Some research questions are not so relevant now but we felt bound to pursue them.

Third, interviewees reported that there was a risk of overlooking unexpected impacts and legacies. Because of this, some suggested that it would have been helpful to revisit and revise the themes and logic models as the Meta-evaluations developed.

The experience of the Meta-evaluation suggests that **defining research questions and logic models at the outset is a useful means of focusing a study on the key impacts and legacies and how these are achieved.** However, they should be **reviewed regularly in the course of a study to take**

account of emerging findings, changing objectives, constraints on data availability and unintended and unanticipated outcomes.

6.5 Data availability

For a variety of reasons, the team found it much more difficult to access data about some of the key impacts and legacies of the Games than they had originally anticipated. In some cases, the secondary evidence from other sources (such evaluations of specific impacts or elements of the Games) was disparate and incomplete. In others, the team had concerns about their reliability. The main problem though, was relevance. This was not because the other studies which the Meta-evaluation drew on were flawed. It was simply that they had collected evidence for a different purpose. They typically focused on one element of the Games, and were often concerned with the delivery of the Games rather than their long-term effects and legacy. One interviewee explained:

> There were reams of data on how the Games were delivered in a sustainable way but nothing on legacy in terms of sustainability.

Some of the studies which the Meta-evaluation drew on were focused on specific localities. For example, there was plenty of evidence about the economic impacts of the Games in London (partly because the former London Development Agency had funded a strong programme of evaluation on employment and skills), but there was much less information about the effects outside the capital:

> 'There were no overarching national evaluations of the impact of the Games on tourism or business or inward investment or trade ………. Modelling of the economic impacts of the Games was based largely on data collected by the Meta-evaluation team.

As a result, the Meta-evaluation team found that it had to spend a lot more time and resource than originally anticipated procuring evidence to fill gaps in the secondary evidence. In some cases they collected additional evidence themselves; often they influenced other studies to collect additional material on their behalf. One interviewee explained that it was important to:

> champion data collection by others and collect primary data ourselves. Primary data was essential to the Meta-evaluation. We had to fill in the bits and pieces of secondary data.

The DCMS played an important role in facilitating this process. It increased the budget for primary data collection by the team and also helped to persuade other agencies (such as Sport England and the Arts Council) to add questions to studies and surveys that would be useful to the Meta-evaluation. It was able to exert considerable influence over the research conducted by related bodies. The Meta-evaluation team reported that the DCMS had also done a good job of encouraging other government departments to take account of the needs of the Meta-evaluation in the studies they commissioned, though its ability to do this was reduced when the client team was reduced.

A number of government funded evaluations which the Meta-evaluation planned to use were scaled back or cancelled after the 2010 General Election including, for example, an evaluation of regeneration in East London. And the Meta-evaluation team experienced difficulties accessing data held by some of the other agencies involved in the delivery of the Games, particularly the London Organising Committee of the Games (LOCOG). Interviewees reported that commercial sponsorship meant there were restrictions on LOCOG's ability to share the data it held, and because it operated at arm's-length from the government the DCMS could not force it to contribute to the Meta-evaluation.

Members of the Meta-evaluation team were asked to comment on surveys and other evidence-gathering tools used by other relevant studies commissioned by other Government departments over the lifetime of the Meta-evaluation. But they were often not able to influence the fundamental design of these evaluations (in many cases considerations of commercial confidentiality meant they were unable to have advance sight of research specifications or invitations to tender), and often their input was restricted to commenting on draft reports and asking questions of evaluation teams after they had conducted their empirical work.

Interviewees reported that, in spite of these challenges, they had managed to compile a lot of evidence, but its quantity and quality varied between themes, and there were some gaps:

> In the end the amount of evidence wasn't bad, but it felt bitty

and left the team with a:

> tricky drafting issue ….. as an author it is uncomfortable because you can see the holes in what you're writing about.

Interviewees suggested that one of the key lessons from their work is that because, by definition, meta-evaluations will be dependent on other studies for evidence, meta-evaluators are not in control of the data available to them, the form it takes or when it becomes available. This means that there is a premium on synchronising the design of meta-evaluations and the studies they draw on. Some of the challenges around data availability can be addressed if **meta-evaluations are commissioned ahead of the studies which they will draw upon**. It is also **helpful if meta-evaluators have an input into the terms of reference of the studies on which they will draw and on-going channels of communication with them**. This is likely to work best where the meta-evaluation is seen to be adding value other studies, as well as vice versa.

Where it is not feasible to commission meta-evaluations ahead of other studies, it would be **advisable to allow a contingency budget to allow for additional work to fill gaps and analyse unanticipated policy developments and outcomes**. It may also be helpful to **make greater use of types of evidence, for example expert judgement might be deployed to complement 'hard' data**.

Finally, **performance management information from public programmes might be useable in more imaginative ways** to help inform both formative summative evaluations; meta-evaluation methods could be valuable both to assess the quality of such data and how best to integrate them into the overall assessments.

6.6 Quality assurance

Our review of the literature demonstrated that quality assurance of secondary evidence is widely seen as a key task for meta-evaluation. The lead team members echoed this. They saw assessing the robustness of the other studies which they drew on as an essential element of their own work. As noted above, the primary issue confronting them was a lack of evidence in respect of some of the themes. As one interviewee explained:

> the idea that you have lots of suitably relevant and robust data on the same thing turned out not to be true in this case.

Another told us:

> We've just not had enough data. We needed every scrap of evidence in practice so it's rarely been a choice between sources.

As a result, the team often did not have a choice between several different sources of evidence on an issue and so it usually tried to use all of the relevant evidence.

However, the team did adopt a systematic approach to assessing the relevance and the rigour of secondary data using a quality assurance tool developed by academic experts. Some saw this as having been an important and very useful means of testing evidence. One reported that:

> It worked really well. It confirmed what we knew intuitively about data quality but it's nice to have this backed up.

But two members of the team argued that the tool had been counterproductive because it had 'sieved out' potentially useful data:

> The whole story didn't ever come across because unless evidence came from an official evaluation it wasn't included ...... (the QA tool) reduced it all to the very few set piece evaluations that exist because every single piece of data needed to stand up in its own right.

The implication of the Meta-evaluation's team experience is that **meta-evaluators need to undertake systematic assessment of the quality of evidence from secondary sources**. These assessments can be used in two ways**. Where there is plenty of evidence, quality assurance identifies the most reliable sources**. **When there is only one source of evidence, quality assurance should be used by meta-evaluators to identify those conclusions which rely on less reliable evidence and should therefore be given a 'health warning'**.

6.7 Aggregation

Like quality assurance, the aggregation of data from diverse sources is widely seen as one of the defining features of meta-evaluation, and the lead members of the Meta-evaluation team expected this to be a major part of their work. One said that they had expected the Meta-evaluation to be based on the:

> same principles as other evaluations but using other people's evaluations.

Some interviewees had expected that this would mean they needed to develop a distinctive set of meta- evaluation methods. In practice, however, aggregation proved much more difficult than they anticipated because of the lack of suitable secondary evidence. As noted above, in some cases there was only one source of evidence about an issue. Sometimes the coverage was partial (for example covering the Olympic boroughs or London but not the rest of the UK). As one interviewee explained:

> There wasn't much overlapping of relevant and reliable data. It was like piecing together the jigsaw, with gaps and with evaluations at different spatial levels, some detailed case studies, some broader.

Sometimes datasets were incompatible because they had been collected for different purposes, in different forms, at different times.

Because of the lack of secondary data, the methods used by the Meta-evaluation team to analyse evidence were often similar to those employed in other large policy or programme evaluations which they had worked on in the past. Rather than bringing together large datasets, or performing new analysis on them, team members described the process as:

> splicing together [separate] arguments and facts from the different pieces of research.

Another explained:

> You're not having to come up with sophisticated ways of aggregating data. It's more about putting together the jigsaw from the data you do have.

Often the team synthesised findings rather than data from other studies:

> We pulled out the key findings in each sub-theme, dropped them into the text, and then looked at the overall story to see if the messages were consistent.

As a result, the Meta-evaluation did not push the methodological boundaries in terms of data aggregation. As one interviewee explained, meta-evaluation methods:

> should be different. However, in practice our meta-evaluation has not been different – it has been very like other evaluations, partly because of the emphasis on primary data collection.

And team members believed that the process of analysis which underpinned the Meta-evaluation of the London 2012 Games was not very different to that of other large-scale evaluations:

> it is probably a fact that all big governmental evaluations have an element of bringing in results from other evaluations.

Our review of other meta-evaluations commissioned by the UK government and studies of other mega-events (chapters 3 and 4) supports this view. Most of the studies we identified had experienced similar challenges and it seems clear that the ability to **aggregate data from diverse sources will be rare**, even in meta-evaluation. **It is made very difficult if there is no effective planning and orchestration of studies. Meta-evaluators could play a role in enabling this but to do so they need to be in place early in the overall evaluation process.**

Another clear pointer from the Meta-evaluation of the 2012 London Games is that in the real world meta-evaluation will often be as much about **synthesis of research findings** as it is about aggregation of data from disparate studies and this is an area in which there would be benefit in **work to develop meta-evaluation skills and tools.**

One specific research sub-theme where aggregation of quantitative data was possible relates to the impacts of the 2012 Games on tourism. Here, data from several surveys was bought together to produce an overall assessment of impact. This was made more feasible by the fact that the team were able to influence and have close control of the design of some of the questions in each survey, which were aligned to fit with the team's impact assessment model.

6.8 Counterfactuals

Another challenge for the Meta-evaluation was the difficulty of constructing reliable counterfactuals against which to measure impact and legacy. The team had planned to establish strong counterfactuals in Report 3 and then report progress against them in Reports 4 and 5. In practice, they were able to develop what one interviewee called 'very good policy counterfactuals' by asking stakeholders about what policies they developed because of the Games. But they lacked hard evidence about what impact there would have been on outcomes if these policies had not been enacted.

There were several good reasons for this. Few of the studies which the Meta-evaluation drew on produced counterfactuals of their own. It was difficult for the Meta-evaluation team to go as far back as 2005 when the Games were awarded to London and construct counterfactuals. And it was very difficult to isolate the effects of the economic crisis in 2008 from the impacts of the Games. Moreover, in many cases there were only a small number of data points, and there were often problems with baselines. At the same time, the Meta-evaluation was too short to be able to measure longer-term legacy effects. So the team had to resort to using evidence which asked respondents to speculate on their longer term decisions and behaviour in order to estimate the impacts of the Games beyond the lifetime of the study.

The strength of counterfactuals varied between themes. They were strongest in the case of the East London theme because here planning documents gave 'Games on' and 'Games off' scenarios, though again the 2008 recession complicated the picture. Most other themes reported overall trends and used qualitative evidence to help estimate counterfactuals. One interviewee reported:

> The evidence was sufficient to be able to say sensible things in most areas but this is only ever part of the story. The evidence was always provisional because of the timing issue and caveats about attribution of impacts so we used higher and lower estimates.

Problems producing robust counterfactuals are not, of course, unique to meta-evaluation. But the scale and complexity of mega-events make it particularly difficulty. **If meta-evaluation is worth doing, then an attempt at constructing counterfactuals is usually going to be worth trying, even though it will, by definition, be difficult and less than perfect.** There are benefits in **starting meta-evaluations early in order to give them the best possible chance to capture baselines**. And it is **important to take the needs of a meta-evaluation into account in the design of other studies**. This in turn **highlights the importance of a coordinated, cross-government approach to the evaluation**.

6.9 Defining legacy

The Meta-evaluation team reported that it was difficult to define precisely what was meant by 'legacy' and in particular which initiatives should be considered part of the attempt to secure the legacy of the London 2012 Games. Some initiatives were entirely new interventions that were directly associated with the Games; others were pre-existing programmes that were 'rebadged' as part of the Olympic/Paralympic strategy. Different legacy strategies were produced at different times, and some (for example those relating to tourism and export growth) were not published until after the Games. And the priority attached to legacy changed in the course of the study. Interviewees said that at the outset the focus was very firmly on the long-term legacy of the Games. Over time, there had been increasing interest from policy makers in more immediate impacts.

The change of government in 2010 also complicated the picture because it was seen as having led to changes of emphasis and priorities, for example a shift towards competitive school sport. The result was that as one interviewee explained:

> Report 1 set out policies to be included but new programmes come along and you can't ignore them so the boundaries get blurred and you end up evaluating vastly more programmes than expected, which detract from 'core' policies, programmes and projects.

The result was that the Meta-evaluation team had to try to 'retrofit' its approach and findings to emerging strategies and priorities. The team said they would have valued a clearer steer from the Government about what it was most interested in, in terms of learning from the Games:

> It is only now giving clues about what it wants to know about in terms of lessons - impact on women, management of mega-events, and volunteering.

Again, this challenge is not unique to meta-evaluation, but the scale of London 2012 Games makes it a more a difficult task than is the case for an evaluation of a single programme or project. The experience of the Meta-evaluation team suggests that it is **important at the outset of a study to**

develop a clear definition of legacy and of the mechanisms for securing it and to try to stick with these throughout a project.

6.10 Managing complexity

What made the study challenging was its combination of scale, scope, complexity, length and political sensitivity. The scale and scope of the Games meant that a wide range of government departments had an interest in its impacts and legacy. As a result, the team had to liaise with a very large number of stakeholders and studies - inside and beyond government. This proved time consuming, though team members reported that DCMS had shouldered much of this burden. The Meta-evaluation was, we were told:

> steered by committee and we needed to keep lots of stakeholders happy …….. we needed to serve all departments' interests.

One of the problems facing the Meta-evaluation was that different stakeholders are interested in different aspects of the study but few are interested in the overall analysis. One members of the team cautioned that:

> in a silo government, people only want to hear about their silo – actually, the big picture is not interesting to anyone.

The scale of the Games meant that writing the reports was a huge undertaking. The final report documents totalled more than 1,000 pages and it was difficult to draw the diverse themes together into what one interviewee described as ‘an interesting overall story’. The size of the report also presented the Meta-evaluation team and DCMS with a considerable quality assurance task. This wasn’t technically difficult work but required:

> late nights spent in the office reading through the draft report.

The study was not only large, it was also relatively long. This meant that there was inevitably turnover of personnel - in departments and in the evaluation team - and the team had to work at maintaining and renewing its links with key stakeholders. At the same time though, the team had worked under considerable time pressure, particularly towards the end of the study. Several important evaluations which they hoped to draw on were not published until the final stages of the Meta-evaluation, which meant they had only a short period in which to incorporate these into the final report:

> We were waiting for the results of other evaluations. Some didn’t arrive until April 2013 so we left gaps in draft reports and had to drop in material at the last minute.

These pressures were compounded by the high profile nature of the Games and the understandable public interest in its impacts and legacy. As an interviewee explained:

> The team has had to run around gathering evidence that has only appeared a few weeks before the reports …….. DCMS was sensible and sympathetic but there were pressures on them too including the fact that the PM wanted to do a statement on the anniversary of Games.

Some of these challenges are unavoidable. However, they highlight **the need for a powerful ‘customer’ for meta-evaluation who is interested in the overall picture it presents**. It might, for example, have been beneficial if the Treasury had taken a more active role in the Meta-evaluation of the London 2012 Games.

6.11 Skills and Capacity

The challenges identified by the Meta-evaluation team were reflected in the skills which they said they believed meta-evaluation called for. They saw the ability to bring diffuse evidence together in one report as a key requirement and believed this made meta-evaluation different from other studies they had conducted. The technical skills to construct logic models were also necessary for meta-evaluation, as was a firm grasp of how to analyse primary and secondary data.

The scale, scope and multi-faceted nature of the work, plus the high profile of the London 2012 Games, meant that effective project and risk management were even more important than in other types of study. The work also called for effective leadership of a large multi-disciplinary team from three organisations and other experts. And the reliance on data from others meant that good stakeholder management was essential. The ability to influence other agencies and other evaluations was paramount:

> The ability to engage with stakeholders is really, really important.
>
> There was a lot more consultation involved than we had expected, needing to find out what other stakeholders were doing around evaluation, and to get them to understand what the Meta-evaluation was about and trying to get their buy in.

The team also spent a lot of time and effort in formal and informal consultations to get qualitative evidence and several interviewees emphasised the important of ‘persistence’ and ‘endurance’. One interviewee concluded that:

> Persistence is a prerequisite.

This suggests that **if research councils and/or others wish to enhance the capacity for meta-evaluation within the UK research community they should pay attention not just to training in methods but also to developing research leadership and stakeholder management skills**.

6.12 Using the Meta-evaluation

The Meta-evaluation team identified a wide range of potential users of and uses for their work, and interviewees were optimistic that its findings would be taken up and all but one of them recommended that other countries staging mega-events (including future Olympic and Paralympic Games) should conduct meta-evaluations of them. The Meta-evaluation was unique in that it attempted to bring together evidence about all of the different facets of the Games to provide the ‘big picture’ and users included UK government departments and local authorities, arm’s-length bodies such as the sports and arts councils, cities and countries staging similar large events in the future, including Glasgow and Rio.

Interviewees pointed to a wide range of useful lessons for policy makers, both in relation to the staging of mega events and a range of broader imperatives. For example, they saw the experience of the Games Makers as being very relevant to the Government’s broader objectives in relation to the ‘Big Society’.

In addition to the lessons for future policy, interviewees believed that the Meta-evaluation also provided an important mechanism for accountability. They said that the scale of the public funding for the London 2012 Games made it important that there was an independent check on whether they delivered the legacy that the Government had anticipated, beyond the ‘feel good’ factor that came from having staged a ‘good Games’. Some interviewees suggested that to ensure complete transparency, it would be desirable to appoint an independent chair of meta-evaluation steering groups.

7: Value to policy makers: lessons from the London 2012 meta-evaluation

7.1. Introduction

The final part of this study was to explore this issue of how the value of meta-evaluations to relevant policy makers can be increased. A workshop was therefore held in August 2013 with a group of central government policy makers and some members of the research team. It specifically covered the impact of the London 2012 meta-evaluation from a policy perspective, and lessons on how the impact of such meta-evaluations might be increased when undertaken in future. The central questions addressed were:

- How could the Meta-Evaluation have been more useful to government (national and local)?
- What have we learnt from the process of conducting the Meta-Evaluation to improve future evaluation processes?

7.2 Methodology

Attendees of the workshop included policy makers and researchers from Department for Culture, Media & Sport, Her Majesty's Treasury, Cabinet Office, Department for Communities and Local Government and Sport England. The workshop mainly used a roundtable format but some short group sessions were used to explore specific topics and provide detailed examples of key points. A summary of the key points from the workshop was circulated to all participants, including some extra questions which had not been fully discussed at the workshop, and responses were incorporated into this analysis. A representative of The Growth Boroughs Unit, who was not able to attend the original workshop, also took part in this subsequent round of discussions.

7.3 Who was the audience for the Meta-Evaluation?

The final report of the London 2012 meta-evaluation was distributed to all Government departments (both analysts and policymakers), relevant partners, some 50+ academics, international Olympic organising committees and some other relevant bodies and individuals.

It was recognised as likely that the ‘component’ evaluations would evoke most interest from the bodies involved in the relevant areas – e.g. the sports participation findings are likely to be of most interest to Sport England, the Cultural Olympiad findings to Arts Council England, etc. Ironically, the Meta-Evaluation may not be so interesting for some of these bodies, since they are, in any case, regularly conducting and monitoring some of the information which is in the meta-evaluation. Indeed, it is partly their data which has been fed into the Meta-Evaluation. Moreover, their policy making is more likely to be influenced by data which is directly and narrowly relevant to their core activities, rather than the evidence on wider impacts which the Meta-Evaluation picks up. Inevitably, one of their main interests in the Meta-Evaluation is to ensure that their own activities in relation to the Games have been positively viewed and reported.

This is, however, only part of the story. There are a number of ways in which the Meta-Evaluation potentially provides value-added for these bodies, by highlighting:

- The effects of some of their activities, where in the past they have not been able to afford the kind of detailed surveys or analysis which the Games have triggered.
- Relationships between their core activities and wider variables such as sustainability, impact on people with disabilities, potential role of volunteering, international reputation, etc.
- Gaps in the data which they have on the cost-effectiveness of their activities (especially where the meta-evaluation was able to collect its own primary data).

Of course, many sections of the Meta-Evaluation findings cover issues for which either there is no national body specifically taking a lead or where several organisations have some role – e.g. sustainability, volunteering, impacts on intended target groups included people with disabilities, or the UK's international profile. The challenge here is two-fold – both to disseminate the relevant findings to these stakeholders and to convince them to consider and take action on these findings. For example, the lessons from the Games Makers and London Ambassadors initiatives for other volunteering programme should be utilised by future major sporting and cultural events, although the mechanisms for ensuring this appear weak. The Meta-Evaluation might also be able to help by recommending ways in which future research could indicate whether the actions are working. It is also likely to have provided added value for bodies who do not regularly conduct large-scale evaluation (e.g. Office for Disability Issues) by covering disability in a range of different contexts.

From the rather narrower perspective of value-for-money, the National Audit Office (NAO) has shown a great interest in the Meta-Evaluation from the start. Indeed, if Government had not initiated an overall evaluation of the legacy of the Games, NAO would probably have been very critical. Its interest has been valuable in publicising the report and encouraging others to read it.

However, there are unlikely to be many other bodies in the UK interested in the overall findings, as the Games were a one-off event. Even the Mega-Events team in DCMS were interested only in certain aspects of the meta-evaluation, in so far as it highlights impacts which other future events might also seek to achieve. This highlights the need to focus dissemination on specific themes which may be relevant to particular groups of stakeholders, e.g. volunteering, international reputation, etc.

Finally, from a UK perspective there is the issue of whether the findings of the Meta-Evaluation can be interpreted to make the case that Government should invest more in such big events (although it is important to maintain a sense of proportionality – the 2012 Games were only a very small project in relation to public expenditure as a whole). Again, there is no obvious forum in which this might be raised – but it is another aspect of the difficulty of finding relevant stakeholders at whom to aim specific parts of the Meta-Evaluation findings.

Perhaps the target group most interested in the overall results of the Meta-Evaluation is the group of hosts of mega-events and the bidding commissions of cities bidding for such events. Moreover, the Meta-Evaluation has also been very good news for the International Olympic Committee (IOC) and International Paralympic Committee (IPC), in that it shows the legacy benefits from holding a Games - it is not surprising that the IOC provided a supportive quote for a DCMS press release on the Meta-Evaluation. This is by far the biggest evaluation ever done on this topic, so DCMS has done something enormously valuable for all these stakeholders (without any possibility of recompense).

7.4 What did policy makers expect from the 'Meta-Evaluation'

None of the workshop participants had been directly involved in a meta-evaluation before but all had been involved in the use of research (often evaluation research) in government policymaking.

Discussion explored recent incidents where research of any kind had played a positive role in influencing decisions in which the participants had been involved. Unsurprisingly, a clear lesson highlighted was that research is more likely to get through to policy makers when written in a way which is direct and immediate, not abstract or broad. Another lesson which emerged was that the desire for evidence is most intense in government during the policy formation period. This is in line with the arguments in recent research on evaluation, especially represented by Michael Quinn Patton (1997), that 'process use' is more important than 'use of substantive findings', since findings tend to have a short half-life, whereas process use teaches policy makers a new way of thinking and learning. From this perspective, the London 2012 meta-evaluation had the advantage of being in place for some years before the Games actually took place, and therefore in a position to influence the process of policy making. Indeed, this tendency for early initiation is in the very nature of a meta-evaluation.

Participants suggested that evidence was frequently used in government not only for learning purposes but also to promote policies which were already being considered or had already been adopted. This was mirrored in the Meta-Evaluation. Moreover, there was a general feeling that the Meta-Evaluation was not actually commissioned originally in order to learn lessons but rather to have available answers to (potentially embarrassing) questions about whether the Games were worth their cost to the public. Indeed, this was perceived as an important rationale in some quarters right up to the Games taking place. However, this naturally changed once the Games were perceived to have been successful.

However, participants warned against the presumption that policy makers expected much at all from the Meta-Evaluation – they suggested that very few people would read all of the recent 'final' meta-evaluation report, and even those closely involved may only read short sections of it. Indeed, a month after publication only a few comments had so far come back from those to whom it was circulated. This again reflected the situation that policy makers only have a relatively thin interface with research and have quite mixed motives for using research evidence.

7.5 Did the London 2012 meta-evaluation ask the right questions for policy makers?

While it was agreed that the original questions posed were right at the time they were developed, it was suggested that the Meta-Evaluation framework was possibly too rigid. As policy developed, it would probably have been better to have had some checkpoints at which some adjustments to the research questions could have been made, either because some new questions had become policy-relevant over time (e.g. around women in sport) or because positive or negative unintended consequences (e.g. benefits from more flexible trade deliveries approach in London) were becoming evident. In practice though, it was felt that such changes would probably only apply to a small minority of research questions, since most had stood the test of time.

There was an argument that the Meta-Evaluation might usefully have started with a much smaller core of questions, to which extra questions could have been added to over time, as they surfaced. This would have encouraged greater realism about what can be achieved even by an avowedly 'overarching' evaluation, which inevitably faces many pressures to be highly ambitious. It would also have eased the problem of effective reporting, which was made especially difficult by the sheer scale of the project, the number of stakeholders involved and the amount of evidence. However, there is a counter-argument that 'legacy' is, by its nature, complex and varied, and likely to give rise to lots of unintended consequences, so the more the scope is narrowed, the more likely something of importance will be lost. Additionally, if fewer stakeholders had been involved, this would have reduced the leverage of the Meta-Evaluation in getting information and other inputs from many government departments, which could have reduced rigour.

The Meta-Evaluation benefited from some built-in mechanisms to challenge its research questions, raising issues such as 'unintended consequences' and wider 2012-related activities not originally planned. However, this was not fully followed through in terms of flexibility in the methodology and making a higher contingency budget available to fund primary research into new questions emerging.

7.6 Did the Meta-Evaluation add value by marshalling the relevant evidence?

The balance between an 'evaluation of evaluations' (which is how many see a 'meta-evaluation) and an overarching evaluation which collects primary evidence must partly rest on the circumstances in which the evaluation is undertaken. In the case of the London 2012 meta-evaluation, a number of important evaluations and survey instruments were unexpectedly terminated quite early on in the life of the meta-evaluation, as a result of the change of government and the subsequent public spending cuts – including the DfE Annual PE and Sport Survey and the CLG Place survey which would have provided analysis for the East London theme.

Consequently, the Meta-Evaluation team had to fill in these gaps itself. In fact, DCMS and the team did well in undertaking primary research, finding a range of ways of funding it without having to call much on the Meta-Evaluation budget itself. It has to be recognised, however, that this was made easier in the case of this meta-evaluation because of its focus on the 'magic' Olympics brand – other meta-evaluations would not necessarily find it so easy to initiate primary research, where some of their expected component evaluations fell by the wayside. Moreover, a number of regular data gathering exercises by partners were prepared to include one-off questions on the effect of the Games – again, a mechanism expedited by the Olympics brand.

7.7 Was the time horizon of the Meta-Evaluation appropriate?

Meta-evaluations tend to deal with complex issues, whose full consequences only become evident over a long time period. This was clearly true in the case of the legacy of the London 2012 Games. The difficulties inherent in getting commitment to future follow-up research, which could address these longer-term impacts, were recognised and some steps were taken to deal with this issue. For a start, some of the evaluations which played a role in the Meta-Evaluation are core activities of key partners, e.g. Sport England, and these will continue, allowing a regular check on whether the legacy targets are being met. Moreover, a number of steps taken during the Meta-Evaluation ought to pay off in future years – e.g. some questions were explicitly introduced into longitudinal surveys of different national agencies which will hopefully produce evidence of impact some years into the future when follow-up questions are asked. In addition, there is a commitment by government to provide an annual update on the legacy of the Games to Parliament, which will require compiling latest data from relevant surveys and evaluations completed since the final Meta-Evaluation report. This is an indication that meta-evaluations can leave their own legacy and enhance the exploration of longer-term impacts.

There are also some questions as to whether the Meta-Evaluation started early enough. Workshop participants (reinforcing the views of the Meta-Evaluation team – see section 6.7) felt that an earlier start to the Meta-Evaluation might have helped to influence more component evaluations to take into account the wider issues in which policy makers were interested. Although there was little evidence on whether this was likely to be successful, it is seen as a key building block for successful meta-evaluations.

There were also some issues around the timing of the release of findings from the Meta-Evaluation. There was continuing nervousness from the policy side of government about the publication of reports of the Meta-Evaluation, reflecting concern that negative messages could detract from the preparation and staging of the Games. It was here that a more independent oversight of the Meta-Evaluation might have been particularly valuable.

Of course, the pressures around 'appropriate publication dates' was probably determined largely by the perceptions of how the Games were going. The reluctance to have much published before the Games was probably predicated on the belief that there might be problems with either the staging of the games or its impact, and the Meta-Evaluation interim findings would not necessarily be reassuring on these issues. Then, after the Games, the widespread perception that they had been a success led to the reverse pressure, for early publication. Had the Games NOT been seen to be a success (e.g. if the weather had been poor, the performance of Team GB disappointing, the logistics poorly managed, etc.), then early publication might have been embarrassing. Just as these issues are now easy to underestimate with hindsight, they are extremely difficult to plan for in advance. In the event, the Interim Report came out at exactly the right time – but that cannot really be credited to acute foresight or careful planning.

7.8 What lessons should be learnt for future similar evaluations?

The perceived value of the London 2012 meta-evaluation flags up the potential for meta-evaluations in other government programmes – something which has been tried relatively rarely in the UK government. Its added value came from adding an extra, broader level of analysis and interpretation to the 'component evaluations' on specific aspects of the Games – and in more favourable circumstances might also have involved influencing the component evaluations in such a way as to make their findings more complementary. These benefits seem relevant to a wide range of government programmes, where traditionally project-based evaluations have been commissioned but the bigger picture has not been brought together.

Given the high profile of the Meta-Evaluation, because of its connection with the Games, and the cross-government nature of the Games, it was essential that the Meta-Evaluation was seen to be independent. The greatest benefit of the Meta-Evaluation to DCMS has been its ‘quality mark’ as a rigorous, independent evaluation. Having a cross-government Steering Group helped, by providing ‘checks and balances’ between government departments. However, it may be that some further mechanisms to ensure independence should be built in to such approaches in future (e.g. having an independent body to oversee all major government evaluations or to provide the chair of the Steering Group). This would follow the precedent set by the development of the UK's Office for Budget Responsibility, which helps ensure the Government's economic and fiscal forecasts are independent of Government departments; and also the precedent of the Regulatory Policy Committee that provides independent scrutiny of regulatory measures introduced by Government departments. The development of the new What Works Centre for Local Economic Growth may provide an approach to ensure the independence of future evaluations. This in turn might increase the potential perceived benefits from meta-evaluations.

Part of the problem for the Steering Group was that there was not initially a strong sense of ownership of the Games in government departments (beyond DCMS), and those who did get involved were not always able to state clearly what the priorities were for their organisations. Some departments didn’t ‘wake up to the Games’ until rather late – certainly long after 2009. The fact that there was never a budget for legacy made it difficult – this could have encouraged more enthusiasm by pump-priming the starting of projects. This lack of ownership was perhaps one of the reasons that the research questions were difficult to narrow down – without partners who are clear about what their priorities are, it is riskier to prune questions which might later turn out to be important. This is consistent with the Meta-Evaluation team’s view that it would have valued a clearer steer from government about what it was most interested in, whereas clues about this only emerged slowly and rather late (see section 6.7). Consequently, the research team tended to treat all the research questions equally. Having a more independent oversight of policy evaluations might drive government departments to take evaluation more seriously, with the potential consequence that focused meta-evaluations would gain in attractiveness, as they would set ‘component’ evaluations in context and provide a sense of proportionality to their findings.

There is clearly still a significant divide between the policy side of government and the research community, so that the flow of information remains sporadic and imperfect in both directions. An example from the London 2012 Games was that the Department for Education would have liked some new research on pupils’ perceptions of the Games or an evaluation of the Get Set programme (the London 2012 education programme ran by LOCOG) but it was not able to find any available large-scale research or to get relevant questions into existing survey instruments. At the same time, the Annual PE and Sport Survey had been discontinued, in spite of the strongly expressed concerns of the research community. Eventually the debate which broke out during the Games in a very public way created a situation where there was pressure on government to provide a policy response very quickly. Of course, there will always be some disconnect between the perspectives of the policy and research communities, especially where policy needs might be moving faster than the research community can sensibly respond. However, a key benefit of having a meta-evaluation team in place should be the earlier identification and response to such emerging research needs.

Interestingly, w*e* understand that the National Audit Office, with the London School of Economics, is undertaking a review of Government evaluations, their approaches and how they are used. We will feed this discussion of the London 2012 meta-evaluation into that review*.*

Appendices

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
James Downe, Steve Martin and Tony Bovaird (2012), "Learning from complex policy evaluations", *Policy and Politics*, 40 (4): 505-523.

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).

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text*, 3rd edition. Thousand Oaks, CA: Sage.

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 (adapted from Torgeson et al.., 2008)

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)

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 guides

Meta-evaluation expert interviews

1. 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.

2. 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)

3. 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)?

4. 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)

5. 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?

6. Follow up

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

Meta-evaluation team interviews

1. Introduction

Thank you for agree to meet me to discuss your experience of undertaking the evaluation of the 2012 Games.

As you know, a small group from Ecorys and academia is trying to help advance meta-evaluation methods by:

- Developing a better understanding of what meta-evaluation is and how it works in practice
- Applying this learning to the meta-evaluation of the 2012 Games to help produce a robust study
- Sharing what we learn about methods from meta-evaluation from 2012 Games with the wider research community to help advance future meta-evaluation practice.

So we would like to hear what you think has worked well, what has been problematic, and what lessons you think can be learned for future evaluations.

Your views with be treated in confidence. We will make sure that the findings from these interviews are reported in a way which means individuals' views are not able to be identified.

2. Your role

1. First, could you describe your role on the meta-evaluation?
2. How much experience did you have of similar work prior to the evaluation?
3. Have you been involved in meta-evaluations before this one? If so, please give some examples.

3. Defining meta-evaluation

1. At the start of your work on the 2012 Games, what did you understand by the term 'meta-evaluation'?
2. Has your understanding of what it means changed in the course of the study? If so, how?
3. Do you think meta-evaluation is different from other forms of evaluation? If so, in what ways?
4. To what extent has your work on the 2012 Games involved:
   a) Evaluating the quality and robustness of other evaluations related to the 2012 Games
   b) Synthesising data from a range of sources to provide an overall picture of the impacts and legacy of the Games
   c) Assessing the purpose and theory of evaluation.

4. Designing Meta-evaluation

Objectives

1. Do you think the objectives of the evaluation were appropriate and achievable?

The Logic Model

1. How well did the logic model work for your theme?
2. What were its strengths and weaknesses?
3. What can be learnt from your experience for future studies?

The themes

1. How well do you think the theme-based approach to the evaluation work?
2. What were its strengths and weaknesses?
3. Would you recommend a similar approach for future evaluations (for example of the Rio 2016 Games)?

Cross-cutting issues

1. How well did the approach to cross-cutting issues work?
2. What were its strengths and weaknesses?
3. Would you recommend a similar approach for future studies?

5. Doing Meta-evaluation

Main challenges

1. What have been the main challenges that you have encountered in your work on the 2012 Games?
2. How have you overcome these?

Quality Assurance

1. How well did the QA tool work in your area of the evaluation?
2. How might it be improved for future studies?

Data Gaps

1. Was there sufficient evidence to enable you to undertake a robust evaluation of your theme?
2. What are the main gaps in the evidence relating to your theme?
3. How have you addressed this?
4. What are the lessons for future studies?

Synthesis

1. What challenges have you faced in bringing together evidence in your theme?
2. How have you addressed this?
3. How have you dealt with contradictory evidence?
4. What are the lessons for future evaluations (for example joint working/co-ordination with other studies)?

6. Skills and Capacity

1. What skills did you and your team need to undertake the evaluation?
2. Are there any areas in which you think skills need to be developed to facilitate this kind of evaluation?

7. Using Meta-evaluation

1. Who do you see as the main users/potential users of the evaluation?
2. Do you think the evaluation has informed policy and practice or will do so in the future? If so how?
3. What lessons about the use of evaluation research can we draw from the 2012 Games?
4. Would you recommend that future events undertake a meta-evaluation? If so why?

Appendix 5: London 2012 meta-evaluation tools and reports

Stage 1: Define scope of the meta-evaluation

The following example logic model **clarifies the theories and assumptions** around how the 2012 Games and its legacy activities will impact on community engagement, attitudes and behaviours across the UK.

Rationale	Objectives	Activity	Outputs	Results	Outcomes/Impacts
Market Failure Interventions focussed on building community cohesion provide improvements in social and human capital which are positive externalities that benefit individuals and society more broadly. *Challenge* There are significant levels of inequality within the UK in terms of educational attainment, employment and income levels; social exclusion and issues of cohesion also exist in some communities. There are varying rates of participation in volunteering and culture, influenced by a range of factors such as age, disability and access to opportunities, and varying levels of uptake of more sustainable behaviours. *Opportunity* The 2012 Games provides a unique opportunity to create a lasting legacy of community benefits (and improved well-being) in London and the rest of the UK. This includes community cohesion, social inclusion, education, learning, building active and more sustainable communities and improved attitudes towards disabled people.	To get people setting up their own Games-inspired activities and more people giving time to their communities. Also to create new volunteering opportunities	Volunteering and Community Action	New volunteering opportunities created Volunteers recruited (including young people and hard to reach groups such as low income, BME and disabled) Volunteers accessing training 2012-inspired community activities held and participants involved	More organisations, groups and people set up community activities/offer volunteering opportunities Increased opportunities to volunteer (including in the staging of the 2012 Games, and more widely in the community, sports and arts sector), especially for hard to reach groups including disabled people More people volunteer their time Volunteers gain accreditation as a result of completing training Volunteers gain non-accredited skills (communication, team working, organisational etc) and softer outcomes (confidence, self-esteem, feelings of social inclusion) Increased awareness of 2012 Games and its legacy amongst volunteers and the general public (and sense of pride and belonging) Development of improved volunteering infrastructure (facilitating the matching of demand and supply of volunteer time), community infrastructure (e.g. new groups sustained), and sustainable networks Increased public visibility of disabled people undertaking positive activities	Increased participation in volunteering and involvement in community activity, especially amongst hard to reach groups including disabled people Increased happiness/ subjective well-being Increased satisfaction with neighbourhoods/local area More cohesive and inclusive communities
	To get more people taking part in cultural activities, including increasing disabled people's participation in culture and removing barriers	Culture	Cultural events, commissions and projects People attending/actively participating in cultural activities (including young people and hard to reach groups such as low income, BME and disabled) Case studies/dissemination outputs celebrating disabled people's arts and cultural achievements	Cultural and creative organisations accessing new commissions/contracts Increased access to cultural opportunities, especially for hard to reach groups including disabled people Increased awareness of 2012 Games and its legacy amongst participants/audiences (and pride and belonging) Increased skills, confidence and self-esteem among participants, including disabled participants Increased aspirations/access to employment opportunities within the cultural sector for participants Increased awareness and appreciation of disabled people's arts and cultural achievements Increased interest in (and demand for) future cultural activity	Increased participation in cultural activity across the UK, including for disabled people Increased happiness/ subjective well-being Increased satisfaction with neighbourhoods/local area More cohesive and inclusive communities Growth of cultural and creative sectors (through creation and safeguarding of jobs and GVA)
	To inspire children and young people to aim higher and achieve better outcomes through initiatives inspired by the 2012 Games and the Olympic and Paralympic values	Engaging Children and Young People	Schools and pupils engaged FE/HE sector institutions engaged Development and sharing of resources Scholarships/mentoring provided to hard to reach young people	Increased interest in school/improved attendance/reduced exclusions amongst participants Higher aspirations and increased commitment to education or employment amongst participants Increased self-esteem and development of other soft skills amongst participants Increased awareness of the 2012 Games and its values amongst participants Increased access to opportunities (such as positive educational/career pathways) for participants Participants entering employment/further education/training	Improved social and economic outcomes for children and young people Improved educational attainment Reduced truancy/ absenteeism Increased participation in sport/culture amongst children and young people More cohesive and inclusive communities
	To encourage people to live more sustainably as a result of 2012 Games-inspired activity	Sustainable Living	People engaged with projects Production of resources, tools and events.	Behavioural change amongst participants resulting in reductions in individual resource and energy use and/or development of more sustainable travel patterns Increased awareness of environmental impacts and how to live more sustainably	Reduced energy and resource use by households Reduced household waste production and increased recycling Increased uptake of walking and cycling
	To influence and change attitudes and perceptions of disabled people among the general public, as well as amongst disabled people themselves	Influencing Attitudes Towards Disabled People	Paralympic Games Coverage Spectators attending Paralympic events Positive media articles about Paralympic activity and the involvement of disabled people in the 2012 Games (e.g. in sport, employment, culture, and volunteering) Case studies/guidance/ dissemination of disabled people's achievements	Increased audiences for Paralympic events (spectators and viewers) Increase in the accuracy and positivity of reflections on disabled people's experiences and achievements in the media Increased awareness of Paralympics, disability sport and other 2012 activities involving disabled people and their achievements	Increased feelings of pride and well-being amongst disabled people Improvement in attitudes towards disability among the general public Reductions in the barriers to participation in society and the economy for disabled people More cohesive and inclusive communities

The following research questions formed part of the **evaluative framework** guiding London 2012 meta-evaluation activity, and specifically in relation to the sub-theme of 'volunteering and social action' within the community engagement theme (as well as in relation to synthesising overall lessons learnt across the theme). The questions were developed based upon the Government's legacy plans, discussions with stakeholders and the relevant component of the logic model. The spatial and temporal scope of each question was also specified.

Question	Spatial Scope	Temporal Scope
Volunteering and community action		
To what extent and how have the 2012 Games resulted in more active, cohesive and successful communities, including through:	Nations, regions and host boroughs	To 2013
• Inspiring more organisations to offer volunteering opportunities and building the capacity of the sector?		
• Inspiring more people (and especially young people and disabled people) to volunteer their time, and tackling the barriers to participation?		
• Inspiring people to set up their own 2012 Games-related activities, which engage people across the UK in the Games?		Post 2013
To what extent have any impacts been sustained, supporting the development of the Big Society?		
Lessons learnt		
What lessons can be learned by host cities and countries about how to maximise the community legacy benefit (including cultural, educational and civic benefits) from mega-events? For example in terms of:	Nations, regions and host boroughs	To 2013 and post 2013
• Developing a brand identity.		
• National co-ordination and communication.		
• Strengthening (national and local) delivery infrastructure, including in communities and schools.		
• Sustaining involvement and cohesion benefits (including amongst disabled people).		

The aim was for the answers to these questions to be synthesized alongside the findings from the other sub-themes of the community engagement theme (culture, engaging children and young people etc) in order to help answer the headline thematic impact question of: '*To what extent and how have the 2012 Games resulted in more active, cohesive and successful communities?*'

The full set of evaluation questions for the London 2012 meta-evaluation, alongside the meta-evaluation methods associated with each set of questions, are specified in the London 2012 meta-evaluation reports *Report 1: Scope, research questions and data strategy*[13] and *Report 2: Methods*[14].

[13] See https://www.gov.uk/government/publications/report-1-scope-research-questions-and-strategy-meta-evaluation-of-the-impacts-and-legacy-of-the-london-2012-olympic-games-and-paralympic-games

[14] See https://www.gov.uk/government/publications/report-2-meta-evaluation-of-the-impacts-and-legacy-of-the-london-2012-olympic-games-and-paralympic-games-april-2011

Stage 2: Identify studies

The following table summarises/clarifies the **information required** to help answer the (predominantly impact focused) research questions relating to the volunteering and community action sub-theme, based upon the inputs, outputs and outcomes specified in the logic model. The table also provides a summary of associated accessibility issues identified during the scoping stage.

Sub-theme	Data sources	Key issues
Volunteering and social action	Monitoring and evaluation data from Games Makers, Inspire mark, Personal Best and LOCOG's community engagement programme Evidence provided by GLA, Youthnet and V Citizenship survey (or alternatively LA resident surveys) Active People and Taking Part surveys Understanding Society survey London 2012 Legacy Research Tracker (sample of 665 disabled people)	Evidence from evaluation of volunteering programmes/projects will be limited, but potential to use LOCOG database to provide further data and undertake participant primary research. Citizenship Survey has been discontinued. Some but not all national survey datasets will be available, disaggregated by disability. There are no plans to reinstate the legacy research tracker, but Taking Part provides an alternative source of data for much of the required content.

The more detailed table below summarises the work undertaken to **screen whether the information identified fits with the information required**, in terms of the meta-evaluation logic model and research questions for the volunteering and community action sub-theme. It also identifies any additional primary research or analysis required, following the **comparison of available information against what is required**.

Further detail, including evidence tables for all sub-themes, can be found in the London 2012 meta-evaluation report *Report 2: Methods*[15].

[15] See https://www.gov.uk/government/publications/report-2-meta-evaluation-of-the-impacts-and-legacy-of-the-london-2012-olympic-games-and-paralympic-games-april-2011

Research questions	Extent covered by programme evaluations	Extent covered by survey/statistical data		Required additional survey questions	Required primary research	Required modelling
		Source	**Key questions/ indicators**			
To what extent and how have the 2012 Games resulted in more active, cohesive and successful communities, including through: • Inspiring more organisations to offer volunteering opportunities and building the capacity of the sector? • Inspiring more people (and especially young people and disabled people) to volunteer their time, and tackling the	Evaluation of Cadbury's Sports v Stripes (Ecorys 2011-12) will provide evidence on community engagement outputs and cohesion outcomes based on participant research using SROI Framework. Youthnet undertakes on-going survey work including a question on whether users were inspired by the Games to volunteer. Also have monitoring data and plan an evaluation possibly using SROI (dependent upon availability of funding post March 2011). Monitoring data from the 'Do It' Website could be used to measure increases in the number of organisations offering opportunities. V will evaluate its Games-related projects as part of a wider on-going evaluation of its work using SROI	Both the Citizenship Survey (since 2001) and Taking Part (since 2005/06) provide a measure of volunteering levels in the general population (England & Wales/England only) and breakdown by disabled people. The Citizenship Survey also provided indicators which can be used as a proxy for community cohesion but will not continue beyond 2010/11. Relevant questions are being considered for inclusion in Taking Part. Active People survey Sport England: “volunteering to support sport for at least one hour a week” available	Taking Part includes a question to explore the influence of the 2012 Games on participation in volunteering (‘*do you think that the UK hosting the 2012 Olympic and Paralympic Games has motivated you to do more voluntary work?*’) Taking Part for 2008-9 and 2010-11 monitors volunteering and health/disability status (not measured in 2009-10). Wave 5 of Active People onwards contains detailed disability question with 10 types of impairment. Survey	Continuation of questions which form a proxy for community cohesion in national surveys (relevant questions are under consideration for inclusion in Taking Part). Further modification of Taking Part to explore nature/extent of volunteering activity influenced by the Games (such a question is currently being tested for potential inclusion from 2011/12). A 2012 Games volunteering-related question will be included in Understanding	LOCOG database offers potential for a participant survey to be administered on behalf of the Meta-Evaluation team (ideally to include a sub-sample of disabled volunteers). Potential for additional research or case studies into benefits of volunteering for people with a disability. Self-evaluation template/exit survey of Inspire projects could be rolled out nationally, via LOCOG, DCMS, or Meta-Evaluation team	Recent trends in participation will be projected forward as part of the process of developing the counterfactual.

barriers to participation? • Inspiring people to set up their own 2012 Games-related activities, which engage people across the UK in the Games?	approach (NATCEN/IVR 2010/11). V can also provide monitoring data. Evaluation plan developed for London Ambassadors. This work is yet to be tendered but expected to be undertaken in 2011/12, to include qualitative research with volunteers and assessment of (longer term) tourism benefits. Inspire monitoring data will provide further evidence of community participation in Games-activity. The GLA plan to adopt a self-evaluation template, which could capture further information about impact. No plans for evaluation of Games Makers (applicant database does provide information on prior participation levels). Data from LOCOG on %/n of disabled volunteers compared with previous Games.	disaggregated by disability. London 2012 Legacy Research. Wave 3, 2009 included a sample of 665 disabled people. There are plans to repeat this survey in the future. Life Opportunities Survey contains questions on barriers to volunteering Q241-243	confirmed for 3 further waves. Q27 of Legacy Research tracks increased volunteering as result of 2012 Games	Society in 2012/13 (although findings will not be available within the timeframe of this study).		

Stage 3: Coding from studies

Accessibility issues relating to secondary data sources, progress in commissioning and implementing additional primary research and any other risks were then continuously monitored during the course of the London 2012 meta-evaluation, to help **manage and map the information through the review process**. This was undertaken using an Excel spreadsheet tracker, with one worksheet per theme, and each row of each worksheet listing the major data sources for the respective theme.

Each data source was given a Red, Amber or Green (RAG) rating depending upon its risk and status, and updated on a monthly basis by each thematic lead of the meta-evaluation. This helped to ensure that the **evidence needs of the London 2012 meta-evaluation were being met** across all themes and sub-themes (and that if necessary contingency plans could be put in place), prior to more detailed coding as part of the quality and relevance appraisal and synthesis of the data.

Stage 4: Quality and relevance appraisal

The table and guidance below provide a summarised version of the research tool developed to help assess the quality and relevance of data sources identified for inclusion in the London 2012 meta-evaluation, and to arrive at an **overall assessment of Weight of Evidence (WoE) in answering the research questions**. This tool incorporates an assessment of: **the general rigour by which the information has been produced** (WoE A); **the relevance of the research design for answering the review questions** (WoE B); and **the relevance of the execution of the design for answering the review questions** (WoE C).

One table was completed for each sub-theme of the meta-evaluation. For each data source, WoE judgements were then used to decide whether: (i) to include the findings in the answer to the (sub-theme) meta evaluation question; (ii) to include the findings but to qualify them in some way, due to weaknesses in quality and/or relevance; or (iii) to exclude findings as not being of sufficient quality and relevance (in practice this was uncommon, since at stage 2 we had already 'screened' the evidence for relevance). The aggregate assessment in the final row also allowed for an overall judgement to be made on the WoE available in support of answering the key research questions for each sub-theme, in terms of the quality and relevance of available data.

Meta-Evaluation QA Tool: Community Engagement theme (sub-theme x)

Source evaluation	WoE A						WoE B	WoE C	WoE A+B+C
	Transparency	Accessibility	Propriety	Accuracy	Specificity	Overall	Purposivity	Utility	Summary
London 2012 project evaluation #1									
London 2012 project evaluation #2									
London 2012 dataset #1									
London 2012 dataset #2									
Sub-theme assessment									

Within each field, both summary statements and scores were provided, based upon the framework provided below.

WoE A: Quality of execution of study

Transparency: Is it open to external scrutiny? Is it easy to tell how the evidence/knowledge was generated? Has the research process been documented?

Scoring: 1 = Low transparency/2 = Medium/3 = High transparency

Accessibility: Is it intelligible? Is the information presented in a way that allows us to readily understand and use it (is it, for example, too dense, technical or ambiguous)? Is reporting clear and coherent?

Scoring: 1 = Low accessibility/2 = Medium/3 = High accessibility

Propriety: Is it legal and ethical? Is there any evidence to suggest that the research was not conducted with due care, the informed consent of stakeholders and within ethical guidelines?

Scoring: 1 = Low conformity with acceptable standards/2 = Medium/3 = High conformity with acceptable standards

Accuracy: Is it well grounded? Are the recommendations and conclusions based on relevant and appropriate data (or are they just asserted with little basis in the research itself)?

Scoring: 1 = Low accuracy/2 = Medium/3 = High accuracy

Specificity: Does it meet source-specific standards associated with evaluations? Including, as appropriate, standards for impact evaluations (e.g. Maryland Scale) and/or standards for qualitative research (e.g. Quality in Qualitative Evaluation: A framework for assessing research evidence)

Scoring: 1 = Low robustness; 2 = Medium; 3 = High robustness

WoE B: Relevance of research design to the review question

Purposivity: Is the design fit for our purposes? Are the research framework, methods, sample design etc relevant to answering the key questions for the meta-evaluation? Is there a close fit with our thematic concepts and logic model/theory of change?

Scoring: 1 = Low relevance/fitness for our purpose/2 = Medium/3 = High relevance/fitness for our purpose

WoE C: Relevance of execution of the study design to the review question

Utility: Are the findings fit for our use? Has the implementation of the research design resulted in relevant and useful information for answering our evaluation questions? Can the information presented be used, or is it incomplete or missing important information? Are the findings generalizable/sufficiently contextualised?

Scoring: 1 = Low usefulness/2 = Medium/3 = High usefulness

WOE A + B + C: Summary judgement

Scoring: 1 = report evidence as low in quality and relevance/2 = report evidence as mixed in quality and relevance/3 = report evidence as high in quality and relevance

Stage 5: Synthesis

The synthesis stage of the London 2012 meta-evaluation is best represented by a series of thematic '*evidence base*' reports[16], produced in support of the main Post-Games evaluation report.

Stage 6: Conclusions and dissemination

The main findings and conclusions of the London 2012 meta-evaluation can be found in the London 2012 meta-evaluation reports *Report 3: baseline and counterfactual*[17], *Report 4: interim evaluation*[18] and *Report 5: Post Games evaluation*[19] and in their accessible summary reports.

[16] These thematic reports can be found at: https://www.gov.uk/government/publications/report-5-post-games-evaluation-meta-evaluation-of-the-impacts-and-legacy-of-the-london-2012-olympic-and-paralympic-games

[17] https://www.gov.uk/government/publications/report-3-baseline-and-counterfactual-meta-evaluation-of-the-impacts-and-legacy-of-the-london-2012-olympic-games-and-paralympic-games

[18] https://www.gov.uk/government/publications/report-4-interim-evaluation-meta-evaluation-of-the-impacts-and-legacy-of-the-london-2012-olympic-games-and-paralympic-games

[19] https://www.gov.uk/government/publications/report-5-post-games-evaluation-meta-evaluation-of-the-impacts-and-legacy-of-the-london-2012-olympic-and-paralympic-games