Developing a Research Agenda for Impact Evaluation in Development

Patricia J. Rogers and Greet Peersman

Abstract This article sets out what would be required to develop a research agenda for impact evaluation. It begins by explaining why it is needed and what process it would involve. It outlines four areas where research is needed – the enabling environment, practice, products and impacts. It reviews the different research methods that can be used to research impact evaluation and argues for particular attention to detailed, theory-informed, mixed-method comparative case studies of the actual processes and impacts of impact evaluation. It explores some examples of research questions that would be valuable to focus on and how they might be addressed. Finally, it makes some suggestions about the process that is needed to create a formal and collaborative research agenda.

1 Introduction

In recent years, there has been heated discussion about the best ways to do impact evaluation, driven in large part by concerns about the consequences of doing it badly – erroneous decisions about which programmes to invest in, and an inability to advocate for ongoing funding for development programmes. Much of this debate has focused on methods and designs for causal attribution, but there are other aspects of impact evaluation that have also been debated vigorously. The irony is that, although impact evaluation is intended to ensure that decisions about practice and policy are informed by evidence, the various arguments about impact evaluation have rarely been based on systematic research.

This article provides a starting point for the development of a formal and collaborative research agenda. It begins by defining why a research agenda is needed and what it would cover. It outlines four areas of impact evaluation where research is needed – enabling environment, practice, products and impacts. It reviews the different methods that can be used to research impact evaluation and argues for particular attention to detailed, theory-informed, mixed-method comparative case studies of the actual processes and impacts of impact evaluation. It explores some examples of research questions that would be valuable to focus on and how they might be addressed – not to provide a definitive review of each topic but to illustrate the scope and approach needed. Finally, it makes some suggestions about the process that is needed to create a research agenda that is not just a wish list or arena for fighting for resources by evaluators, but a productive collaboration among the various parties needed to bring the research agenda to life.

2 Why do we need research on impact evaluation in development?

Sometimes, 'impact evaluation' and 'development' are understood narrowly. We argue that a broad description of both is needed.

2.1 *Defining impact evaluation and development*

As in many areas of evaluation, there are different definitions and conceptualisations of impact evaluation. Some definitions restrict impact evaluation to studies which use particular designs – for example, the United States Agency for International Development (USAID) Evaluation Policy defines impact evaluation as involving a constructed counterfactual such as a control or comparison group:

> Impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control

IDS Bulletin Volume 45 Number 6 November 2014 © 2014 The Authors. *IDS Bulletin* © 2014 Institute of Development Studies
Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA

> for factors other than the intervention that might account for the observed change (USAID 2011: 1).

In this article, impact evaluation covers any evaluation which assesses actual or likely impacts – the Organisation for Economic Co-operation and Development-Development Assistance Committee (OECD-DAC) defines impacts as 'positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended' (OECD-DAC 2010: 24). This implies that an impact evaluation has to address longer term results, but not necessarily directly; it could use other data to make links to likely longer term results, and include *ex ante* and *ex post facto* impact evaluation. What is particular about impact evaluation is that it seeks causal inference, understanding the role of particular interventions in producing change. This essential characteristic of impact evaluation determines its importance as a public good in terms of producing evidence of 'What works?' and 'What works for whom in what contexts?'

By development, we are referring not only to projects funded by international aid but also to programmes, projects, policies and strategies that are funded through various means with the aim of improving health and welfare.

Some of the conceptual maps of impact evaluation in development have only included particular types of development, particular types of impact evaluation, and particular aspects of impact evaluations. Much of the discussion has focused on causal inference methods in experimental or quasi-experimental impact evaluation of discrete donor-funded aid projects in order to inform decisions about scaling up interventions that have been found to be effective.

A research agenda on impact evaluation needs to include the larger map of development – not just donor-funded projects, but country-led programmes and policies, public–private partnership projects and civil society development interventions. Consequently, it needs to include impact evaluations for a range of different users – donors and national governments are important, but also the decentralised level of government responsible for implementation, non-governmental organisations (NGOs) and the private sector who are also engaged in delivery. Last but not least, it needs to include impact evaluations that treat communities as users of evaluations, drivers of development effectiveness and agents of evaluation themselves.

The research agenda needs to include all scales of intervention – not only individual projects, but also programmes, multiple projects as part of a single programme, strategies and policies. It should also include impact evaluations that look at when particular intervention types are suitable – projects which aim to catalyse or coordinate, impact investment, pay-for-performance, or capacity development.

2.2 *The need for a research agenda on impact evaluation*

Impact evaluation can make an important contribution to development. The results can inform decisions about what to invest in, and in what situations, and how to adapt successful projects, programmes and policies for new situations. Evidence of effective development interventions can be used to advocate for continuing or increased funding, especially in a climate of increasing scepticism about the value of international aid. The process of impact evaluation can improve communication between stakeholders and focus attention on results. It can support the principles of effective development, including partnerships and local agency. But impact evaluation can also harm development. Poor quality impact evaluation (either using methods and processes poorly, or using inappropriate ones) can provide invalid, misleading or overly simplified findings. These can lead to poor decisions, such as scaling up interventions that are ineffective or harmful, or that are implemented in situations where they don't have a chance to work. Poor quality impact evaluation processes can undermine developmental processes, reinforcing power disparities and reducing accountability to communities.

Concerns about the impact of poor quality impact evaluation have led to vigorous and sometimes vitriolic debates about appropriate methods for impact evaluation. For example, at a symposium on evidence-based policy, the alternatives to experimental and quasi-experimental designs were summarised as performance measures, customer satisfaction and 'charlatans' (Smith and Sweetman 2009: 85).

However, recommendations for practice have rarely been based on systematic and empirical evidence. It is difficult to secure funding for research into evaluation, and there are few incentives for organisations to collaborate on the sorts of research that would be needed. An international research agenda for impact evaluation would help to build a much needed evidence base for more effective and appropriate impact evaluation. The research agenda could provide a focus for research and an impetus and incentive for joint research across the various sectors, disciplines and organisations involved in impact evaluation of development. It would help to secure commitment and resources for research and to prioritise where these might be applied best. It could support agreements about priority areas for research and appropriate methods for doing this research, and help to make better use of the research that is done by supporting synthesis and dissemination.

3 What is a research agenda and how should it be developed?

To be effective, a research agenda cannot simply be a wish list developed by researchers, nor an ambit claim developed by a self-selected group. It needs to be inclusive, transparent and defensible. To maximise uptake of the findings, it needs to encompass strategies and processes for engaging intended end users in the research process, including in the process of identifying and deciding research priorities.

Some recent examples from the public health arena may provide useful insights in terms of what is needed to get to a research agenda on impact evaluation in development (see, for example, MOHSS/DSP 2010; NAMc 2010; Peersman 2010). In response to a call for an increased focus on programme evaluation to improve national HIV responses, the Joint United Nations Programme on HIV/AIDS (UNAIDS) supported governments[1] in the implementation of a *national evaluation agenda for HIV*. The first step in the process was to develop a national evaluation strategy describing the rationale and objectives for targeted programme evaluations and the procedures and infrastructure for coordination and management of the studies. Formal agreements build on existing roles and responsibilities rather than setting up parallel systems and capitalise on the comparative strengths of different organisations involved. A transparent, standards-based and consultative process was then used to identify key information gaps and to prioritise evaluation studies.

Bringing users of evaluation findings and evaluators together helped to ensure that selected studies were pertinent to the decision-making needs within the national AIDS programme (at all implementation levels) rather than just serving the needs of research institutions, evaluators or funders. It also helped to identify where common interests could be galvanised and unnecessary duplication avoided. There was also more synergy between new and completed evaluation studies and a greater willingness to share evaluation findings. A clear rationale and a costed plan for the implementation of prioritised studies helped to mobilise the funding needed (NAMc 2010).

Understanding what was already known (and thus, where important information gaps exist) was an essential preparatory step in helping to decide evaluation priorities. However, it proved a time-consuming and challenging task as the information was often scattered and not always available in the public domain. Hence, sufficient resources and time need to be provided to do this step well.

Involving a range of different stakeholders with different interests, understandings and/or capacities for evaluation required consensus-building as well as capacity development. These additional efforts allowed for the perspectives of different stakeholders to be heard and appropriately accommodated. It was particularly important to conduct the prioritisation of evaluation studies in a transparent manner and according to agreed criteria such as, for example, the following considerations:

1 The study needs to address an *important data gap* for *improving* the national AIDS programme:
important – the potential for impact of the findings is high; addresses 'need to know' not 'nice to know'
data gap – the question cannot already be answered by existing studies, available data or information
programme improvement – the evaluation provides information on what can be done better in terms of programme implementation, effectiveness, and/or efficiency;

2 The study addresses an *immediate need* – i.e. it provides timely data needed for key decisions in the next five years;
3 A *good quality* study is *feasible* (time frame, capacity, cost).

The consensus-driven process was facilitated by an honest broker, someone with both evaluation and facilitation experience and who did not have a stake in what was being prioritised by whom. A first cut at prioritisation was achieved by facilitated discussions in small multi-stakeholder groups, the results of which were consolidated in plenary discussion.

Although the prioritisation of studies focused on addressing important information gaps in the short to medium term, the institutionalisation of the procedures and infrastructure allowed for the process to be repeated and address new information needs over time.

4 What needs to be researched?

Research is needed into four different aspects of impact evaluation: the enabling environment (policies, guidelines, guidance, formal and informal requirements and resources); practice (how impact evaluation is actually undertaken); products (the reports and other artefacts produced by impact evaluations); and the impacts of impact evaluation, including intended uses and other impacts. Some research will focus on only one of these but particularly useful research would link these, building evidence that could be used to develop contingent recommendations about the types of enabling environment, practices and products that are likely to produce beneficial impacts, and how to achieve them.

Across these different areas, different types of research are needed. *Descriptive research* would document what is being done, developing typologies and identifying patterns. *Causal research* would identify the factors that produce these patterns. *Evaluative research* would compare the actual performance to explicit standards of performance. A single research project might encompass more than one type of research. For example, a study of the guidelines that support and direct impact evaluation within development organisations could include descriptive research that documented the different types of guidance provided across different organisations, analysed to produce a typology in terms of the research designs that are acceptable or encouraged. It could include causal research that identified the factors that produced these variations across organisations. And, it could include evaluative research that compared these guidances to quality standards and made judgements about their quality.

Evaluation is not a technology that can be simply imported to new areas of application, but a practice that needs to be undertaken in ways that suit what is being evaluated and the situation of the evaluation. This means that we would need detailed descriptions of the context and what is done as well as what the consequences were. In addition, contingent advice needs to be developed for what methods and processes to use for particular situations, and how to support good impact evaluation.

These different types of research are discussed in more detail in the following sections, and illustrative examples of research questions, some focused on an individual impact evaluation and some on more than one evaluation, are included in Table 1.

4.1 *The enabling environment for impact evaluation*

Individual impact evaluations operate within a larger context of local guidance, policy, capacity development and formal and informal incentives. However, these are not always available for external scrutiny. Research could document the variations and develop typologies of different types used. This would be useful as a resource for other organisations to use and adapt, rather than re-invent the wheel. They would also be useful to combine with research into practice, products and impacts to develop knowledge about what types of guidance and enabling environment are effective in supporting quality impact evaluation – and the extent to which this varies depending on the organisational context and the nature of the development intervention being evaluated.

An example of this type of research was a study of guidance for the development and use of logic models and logframes among development organisations (Wildschut 2014). Manuals and guidelines from different bilateral and multilateral development agencies and international NGOs were compared in terms of the definitions used and the nature of the logic models used. The research found 120 different versions of logic models, which could be grouped into four broad

Table 1 **Types of research into impact evaluation with illustrative research questions**

	Descriptive – what does it look like?	**Causal** – what are the factors that make it like this?	**Evaluative** – in what ways and to what extent is it good?
Enabling environment – guidance, requirements, policies, formal procedures, requirements and expectations	How is impact evaluation defined in official guidance?	What factors influence how prescriptive guidelines are? What formal and informal incentives and disincentives exist for conducting and using impact evaluation?	To what extent do guidelines provide technically correct advice and prescriptions for evaluators and evaluation commissioners and managers?
Practice – what is done in an evaluation	To what extent are impact evaluations conducted in accordance with guidelines? What are the strategies used to elicit and use the values of intended beneficiaries in planning and undertaking the impact evaluation? What techniques are used when baseline data are not available? How is process tracing used for causal inference when a counterfactual cannot be constructed?	What factors influence the level of involvement of intended beneficiaries in impact evaluation decisions and processes? What factors influence or facilitate the use of process tracing in impact evaluations?	How effectively do impact evaluations incorporate the values of intended beneficiaries? How valid are reconstructed baselines? How credible are causal inferences made on the basis of process tracing?
Products – reports and other documents produced during an evaluation	To what extent are evaluation reports consistent with guidelines? What methods of data visualisation are used to communicate findings?	What factors influence full disclosure of technical limitations of impact evaluations? Does a focus on reporting and data visualisation lead to more or less attention on the quality of data collection and analysis?	How validly do evaluation reports present findings?
Impact – influence of report and process on decisions, actions and attitudes	What are the intended and unintended impacts of impact evaluation reports and processes?	Under what conditions does the involvement of intended users in the impact evaluation process produce higher engagement and use?	How can evaluation contribute to social betterment?
Combined	Under what conditions are external evaluation teams seen as more credible than an internal team or a hybrid team?	Do narrow definitions of impact evaluation (constructed counterfactual) lead to lower investment in interventions where this design is not possible? Do simple messages of average findings produce more or less engagement and support among decision-makers?	To what extent do impact evaluation policies affect what can be evaluated?

Source Authors' own.

Table 2 **Key evaluation tasks organised in seven clusters**

Cluster of impact evaluation tasks	Specific tasks
1 **Manage** an evaluation or evaluation system	– Decide what is to be evaluated – Understand and engage stakeholders – Establish decision-making processes for the evaluation – Decide who will conduct the evaluation – generally (external, internal, hybrid) and specifically (choosing an evaluation team) – Determine and secure resources – Define ethical and quality evaluation standards – Document management processes and agreements (e.g. Request for Proposal, contract) – Develop planning documents for the evaluation (e.g. evaluation design, work plan) – Review evaluation (do meta-evaluation) – Develop evaluation capacity
2 **Define** what is to be evaluated	– Develop initial description – Develop programme theory/logic model – Identify potential unintended results
3 **Frame** the boundaries for an evaluation	– Identify primary intended users – Decide purpose – Specify the key evaluation questions – Determine what 'success' looks like
4 **Describe** activities, outcomes, impacts and context	– Sample – Use measures, indicators or metrics – Collect and/or retrieve data – Manage data – Combine qualitative and quantitative data – Analyse data – Visualise data – Generalise findings
5 **Understand** causes of outcomes and impacts	– Check the results support causal inference – Compare results to the counterfactual – Investigate possible alternative explanations
6 **Synthesise** data from one or more evaluations	– Synthesise data from a single evaluation – Synthesise data across evaluations
7 **Report and support use** of findings	– Identify reporting requirements – Develop reporting media – Develop recommendations – Support use

Source BetterEvaluation.[2]

types. In some cases, reasons for the variation were explained in the documentation.

The enabling environment includes both formal and informal processes, and not all of it will be visible in formal documentation. Some of it will be in the form of verbal explanations of 'the way things are done here'. This has implications for the research methods needed to study the enabling environment, which are discussed in Section 5.

For example, Coryn *et al.* (2007) reviewed the models and mechanisms for evaluating government-funded research. They examined the processes used in 16 countries where there were sufficient data to undertake the analysis, and developed a typology of models. A purposive sample of judges rated each of the models in terms of 25 indicators related to five criteria: validity, utility, credibility, cost-effectiveness (monetary and non-monetary), and ethicality. Scores were then weighted and reported.

4.2 *Impact evaluation practice*

While many discussions about impact evaluation have focused only on designs for causal inference, impact evaluation practice involves much more than this. It involves up-front work by commissioners of evaluation to decide what should be the focus of an impact evaluation and how it should be managed. It involves other tasks during the actual evaluation, including selecting appropriate measures and negotiating the criteria, standards and weighting that will be used to make evaluative judgements (especially if the impact evaluation is intended to provide a comparison of alternatives). And, it involves activities after an evaluation report is produced, including dissemination and support to use the findings, meta-evaluation, and, in many cases synthesis of findings from multiple evaluations. It is helpful to think about this broad scope of impact evaluation in terms of seven clusters of evaluation tasks (see Table 2).

4.3 *Impact evaluation products*

Evaluation reports are just one of the products produced by an impact evaluation. Important artefacts are produced at the beginning of the process which may include: the rationale for undertaking an impact evaluation of this intervention at this time; the terms of reference, scope of work or request for proposal produced to brief potential evaluators; the proposals they develop in response, often outlining a design for the evaluation; an inception report developed as a first deliverable, sometimes including a revised design. During the evaluation, interim and progress reports are produced and at the end, in addition to a final report, there can be policy briefs, briefing notes, audiovisual versions of the report and social media reporting.

Documenting, describing and analysing these products would not be easy, since many would be internal documents or subject to commercial-in-confidence restrictions. Overcoming these barriers would provide useful evidence of the different formats and contents of these products as well as evidence of their quality. It would be particularly useful to undertake research which looked at the products and the processes used to produce them.

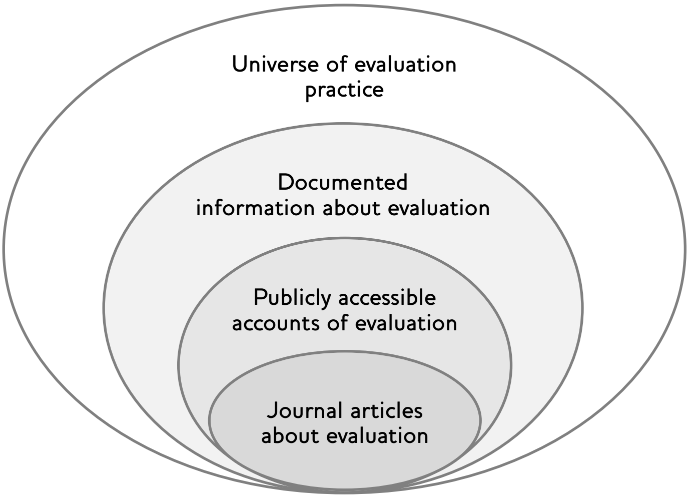
4.4 *The impacts of impact evaluation*

The intended and actual impacts of impact evaluation are an essential element of research on impact evaluation. This research needs to address the different ways in which impact evaluation is intended to be used. Some impact evaluation which aims to discover 'what works' is intended to inform decisions about which interventions to scale up. There is now more interest in learning 'what works for whom in which circumstances', using either realist evaluation (Pawson and Tilley 1997) and realist synthesis (Pawson 2002), or more differentiated experimental or quasi-experimental designs (White 2009). For development interventions which will work differently in different situations (and this would include many if not most of them), impact evaluation needs to also inform users how to translate an intervention to other settings, making appropriate adjustments, not simply transferring it.

Research into the intended and actual impacts of impact evaluation needs to be informed by previous research on evaluation use and influence, including the extensive research on evaluation utilisation (e.g. Patton *et al.* 1975; Cousins and Leithwood 1986; Shulha and Cousins 1997) and more recent research into different ways in which evaluation can and does have an impact (e.g. Valovirta 2002; Mark and Henry 2004). It should also take account of the ways in which impact evaluation can influence the different types of processes involved in implementation, including formal decisions, policies and processes, street-level bureaucrat 'workarounds', devolved decision-making in small groups, conflict and bargaining and responding to chance and chaos (as outlined by Elmore 1978, and its implications for evaluation practice and research explored by Rogers and Hough 1995).

The actual impacts of impact evaluation are not, however, always positive, and research needs to be able to explore unintended negative impacts such as loss of trust (e.g. Schwarz and Struhkamp 2007), the damage done to communities through intrusive, time-consuming data extraction (e.g. Jayawickrama 2013), goal displacement and data corruption in situations of high stakes evaluation (Perrin 2002).

Figure 1 **Journal articles as a sub-set of the universe of evaluation practice**


Universe of evaluation practice
Documented information about evaluation
Publicly accessible accounts of evaluation
Journal articles about evaluation

Source Authors' own.

5 How impact evaluation should be researched

5.1 Options for collecting and analysing data about impact evaluation

A research agenda would not only focus attention on specific research topics but also on the types of research that are needed to investigate impact evaluation for development in terms of its enabling environment, practice, products and impacts. There are a range of methods that should be used beyond those commonly used in published research – surveys of evaluators and analysis of published journal articles about evaluation. The Elmore (1978)/Rogers and Hough (1995) framework raises particular issues about their suitability. If evaluators and programme staff are acting as 'street-level bureaucrats', then they might well be reluctant to disclose their non-compliance with official processes and requirements. If 'conflict and bargaining' processes are important, where different parts of an organisation are engaged in conflict over scarce resources, and collaboration is about temporary advantage rather than long-term commitment to shared goals, then their assessments of the success or failure of the evaluation will be filtered through these perspectives, and probably not willingly disclosed.

The advice from Douglas (1976) is appropriate here. He reminds us that it may be unwise to analyse data as if all respondents are only trying to communicate their perfect knowledge of a situation to the researcher. Our informants' information, and their interpretation of that information, can be affected by their lack of knowledge, lack of self-awareness, and in some cases, deliberate deception.

While Douglas' research was into various forms of deviance, he has argued convincingly that similar issues arise in research into more conventional organisations:

> The researcher can expect that in certain settings, the members will misinform him, evade him, lie to him. This would be true in organised, ostensibly rationalised settings, like bureaucracies. And it is precisely those who are most knowledgeable about these kinds of problems, the managers and the organisational entrepreneurs, who will do most to keep him from learning about the conflicts, contradictions, inconsistencies, gaps and uncertainties. The reason for this is simply that they are the ones responsible for making things rational, organised, scientific, and legal' (Douglas 1976: 91–2).

Research into impact evaluation needs to learn from research into other complex organisational phenomena and use a combination of methods, with a particular emphasis on theory-informed case studies. The following sections outline some of the types of research methods that can be used and issues to be addressed when choosing and using them to study impact evaluation.

5.2 *Literature reviews and systematic reviews*

Published materials can be a useful source of evidence about impact evaluation but care is needed in both data collection and analysis. Formal documents are most appropriate for providing evidence about formal processes, such as guidance and systems. There is still a need to ensure that documents are representative. For example, Wildschut's (2014) review of guidance on logic models and logframes involved an exhaustive search for grey literature.

In recent years, there have been a number of studies which have been labelled as a systematic review of some aspects of evaluation practice but which have been severely limited in their scope – only including journal articles and books – but then drawing conclusions about the state of evaluation practice. For example, Coryn *et al.*(2011) claimed to have conducted a systematic review of theory-driven evaluation, but, for reasons not explained or justified, restricted the search to 'traditional, mainstream scholarly outlets including journals and books', 'excluding other sources including doctoral dissertations, technical reports, background papers, white papers, and conference presentations and proceedings' (p. 208). Journal articles represent a small and unrepresentative sample of evaluation practice, as illustrated in Figure 1.

An earlier comparison of theory-based evaluations reported in academic journals and books with those reported in conference presentations and evaluation reports (Rogers *et al.* 2000) had demonstrated how these are often quite different – the former dominated by academics and by successful evaluations, and the latter by evaluation consultants and including more descriptions of unsuccessful evaluations. Ignoring the larger group omits many of the larger evaluations and a wide diversity of practice and risks making claims about the state of evaluation practice that are not accurate and/or missing the opportunity to learn from more detailed accounts than are published in academic journals.

5.3 *Surveys of evaluators*

Studying evaluation practice and impact through surveys has superficial appeal, especially to graduate students or other researchers seeking to quickly produce findings, but have serious problems. Many surveys of evaluators or organisations have been based on such low response rates that there is real concern about their suitability to provide a picture of the state of practice. For example, the Innovation Network's study of evaluation practice and capacity in the non-profit sector (Reed and Moriaru 2010) reported over 1,000 survey responses but used a volunteer sample with a response rate of 2.97 per cent and a profile of organisations very different to the target population. Even where response rate is not a problem, there remains the challenges of self-disclosure and self-awareness, as well as the question as to whether the person who completes the survey is able to speak on behalf of the organisation.

5.4 *Conference presentations*

Conference presentations can be highly variable as sources of evidence about evaluation practice and impacts. What is very often presented at evaluation conferences and in evaluation journals are descriptions of one's own practice based on poor documentation and in an environment where there are significant incentives to appear competent, minimise the problems, and to make things look neater than the real messy process. There can be barriers to disclosure at professional meetings where people are also seeking employment and engagement.

Some evaluation conferences have, however, managed to provide an environment where people admit challenges and gaps. For example, the American Evaluation Association had a series of sessions on 'Stories of Evaluation' (e.g. Scriven and Rogers 1995) where people were encouraged to share stories of practice. These could be further developed along the lines of the different types of ethnographic stories outlined by Van Maanen (2011): realist; confessional; impressionist; and critical.

5.5 *Simulations and experiments*

Sometimes, it is possible to construct formal tests of different evaluation practices.

Simulation studies have difficulty simulating the group context within which programmes are implemented; most simulation studies on evaluation utilisation have therefore focused on the stage where an individual processes the evaluation information (e.g. Braskamp, Brown and Newman 1982) or makes a decision. It is

Table 3 **Questions about how impact evaluation is undertaken**

Aspect of impact evaluation	Research question	Possible research approach
Develop or use appropriate measures and indicators	What are adequate indicators of important variables which cannot be actually measured?	Review of indicators in use and the understanding of the situation by those using the indicators; peer review of those using the indicators; prize for best indicator in terms of utility and feasibility for a particular challenging outcome
Develop programme theory	How can a theory of change/logic model usefully represent complex aspects (uncertainty, emergence)?	Identification of examples in actual evaluation reports, documentation of process to develop them and usefulness; competition to develop new types of logic models for specific scenarios
	How can an organisation support projects to have locally specific theories of change/programme theory that are still coherent across a programme, an organisation or a sector?	As above
Identify potential unintended results	What are effective strategies for identifying potential unintended results – and for negotiating their inclusion in the scope of the evaluation?	Trials of negative programme theory[3] methods with concurrent documentation of micro-detail of facilitation

Source Authors' own.

difficult for simulation studies to provide enough context for people to enter into a realistic process – and people may respond differently when they actually have an important stake in the evaluation.

More recently, a randomised controlled trial (RCT) was undertaken to test the effectiveness of different types of policy briefs (Beynon *et al*. 2012). A large volunteer sample from various networks on policy and evidence was randomly assigned to receive one of three different versions of a policy brief, and their responses were investigated through an online questionnaire, supplemented by telephone interviews with a sub-sample of each group.

5.6 *Systematic, rich case studies*

Systematic case studies seem likely to provide the most useful evidence about impact evaluation in terms of enabling environment, practice, products and impact – and how these are linked. Different types of case studies (GAO 1990) would be useful. An *illustrative* case study would be descriptive, providing in-depth examples. This could be very useful as a guide for practice, or to develop a typology of practice. Purposeful sampling of the case, including selecting a typical case, would be appropriate. An *exploratory* case study is designed to generate hypotheses for later investigation. Particularly successful or problematic evaluations might be a rich source of new ideas about barriers and enablers to good practice or impacts. A *critical instance* case study might focus on a particular evaluation, or even a particular event or site or interaction within an evaluation which provides a single instance of unique interest (for example, documenting an innovation) or serves as a critical test of an assertion (for example, demonstrating that something is possible). A *programme implementation* case study would examine how evaluation is being implemented, particularly in relation to internal or external standards. A *programme effects* case study would use systematic non-experimental causal inference

Table 4 **Questions about the impact of particular impact evaluation methods**

Aspect of impact evaluation	Research question	Possible research approach
Identify appropriate measures and indicators	Does an emphasis on algorithmic interpretation of evidence-based policy (in the form of identifying 'what works' in terms of a single metric and then ranking alternatives in magnitude of effectiveness or cost-effectiveness) lead to less consideration of equity issues and the implementation of interventions that reduce equity?	Interviews with decision-makers using evidence in the form of a single metric to explore the level of their attention to issues of heterogeneity and equity
Collect or curate data	What are the conditions when big data can provide useful information for impact evaluation?	Review of existing examples of big data use to develop a typology of conditions; trials of using big data on a specific challenge
Understand causal attribution and contribution	How can systematic non-experimental strategies for causal inference be used and communicated effectively?	Identify, document and review existing examples; trial approaches with input from research methods specialists working with evaluators
	Does a requirement for constructed counterfactuals in impact evaluation lead to less investment in system-level interventions?	Interviews with evaluation users

Source Authors' own.

methods, such as process tracing or comparative case study, to draw conclusions about what had produced the observed practices, products and impacts of the evaluations.

Documenting what is done and what happens can use a mix of anthropological methods and cross-case comparisons. Despite the concerns about the limitations of self-reported practice, documented practice will be an important source of knowledge. This would include reviewing existing reports of practice and creating new documentation. This could proceed retrospectively – identifying good practice that has happened and reconstructing what happened and why. It could involve concurrent documentation – identifying particular challenges and following alongside different ways of addressing them. It could involve documenting the processes used and the micro-interactions within an evaluation team and with other stakeholders. Research into the practices of highly skilled evaluators and evaluation managers could develop examples and eventually typologies of strategies used to effectively undertake each of the many tasks (previously outlined in the seven clusters) involved in an impact evaluation.

The process of identifying good examples to document and analyse for case studies could include the winning evaluations of awards and prizes offered by various evaluation associations which have been seen to have been of high quality. Another possible research method for identifying and investigating successful cases would be positive deviance (Richard, Sternin and Sternin 2010), which involves identifying rare cases of success and investigating what they are doing differently. What is particular is that the people involved in doing that enquiry are the people who want to use that knowledge. This might be evaluators, seeking to learn from other evaluators who have conducted good impact evaluations, despite challenges, or it might be evaluation commissioners and users seeking to learn from other commissioners and users.

Cases of low-quality impact evaluation, which could provide useful illustrations of problems and/or unsuccessful strategies, might be identified through crowd sourcing. For example, an enquiry to the discussion list XCeval asking for examples of 'ineffective evaluations' prompted 14 responses and candid discussion of

Table 5 **Questions about impact evaluation politics and infrastructure**

Aspect of impact evaluation	Research question	Possible research approach
Choose what to evaluate	What investments and activities are the subjects of evaluation? On the basis of what evidence are decisions made about other investments and activities?	Review of formal records of impact evaluations (where available); survey of evaluators
	What opportunities exist for funding public interest impact evaluation rather than only funder-controlled evaluation?	Review of public interest research examples
Develop key evaluation questions	What are effective processes to identify potential key evaluation questions and prioritise these to a feasible list?	Detailed documentation and analysis of meeting processes to negotiate questions
Supporting use	How can the utility of an evaluation be preserved when the primary intended users leave before it is completed?	Identify, document and analyse existing examples
Reporting findings	How can reports communicate clearly without oversimplifying the findings, especially when there are important differences in results?	User testing of alternative reports
Manage evaluation	What procurement processes support effective selection and engagement of an evaluation team and effective management of the evaluation project?	Interview evaluation managers, evaluators, and contract managers about their processes, the impact of these and the rationale for them, develop a typology of issues and options
Evaluation capacity development	Why does so much evaluation fail to be informed by what is known about effective evaluation?	Interview with evaluation managers about their knowledge and sources of knowledge about evaluation management

Source Authors' own.

the problems in these evaluations (Luo and Liu 2014). For case studies of weak impact evaluations, de-identification might well be necessary.

A writeshop process, either face-to-face or virtual, can be one way to support retrospective documentation and development of detailed case studies. This involves a combination of writing by one or more people associated with an evaluation (often the evaluation team but in some cases the commissioner as well) with structured editing and peer review. Such writeshops can provide a structure for the cases which examine and articulate aspects of their practice they had not previously thought of, and were certainly not reported in the methodology section of an evaluation report (for example, Oakden 2013 provided a detailed account of using rubrics in the evaluation of school leadership; Cranston, Beynon and Rowley 2013 described an evaluation from the different perspectives of the evaluator and the evaluation commissioner).

5.7 *Trials of methods*

Formal trials of new methods, designs or processes would be an important type of research to support. This would require identifying either a promising method and finding a potentially appropriate situation to use it – or identifying a common challenge and finding a combination of methods or processes to address it. This could take the form of a trial where skilled users of methods apply them to a specific evaluation, with not only documentation but also follow-up evaluation of the validity, utility, feasibility and propriety of the evaluation (e.g. Rogers and McDonald 2001).

This research could be undertaken through a call for proposals (where a specific trial is proposed), through a matching process (where an evaluation site and a methodologist are paired up where supply and demand match), or through a competitive approach to research and development where applicants produce proposals which are competitively assessed – with the prize including actual implementation of the plan.

These trials could include longitudinal studies of the use and impact of evaluation, systematically investigating the extent and nature of impact from the findings and the processes of evaluation.

If one of the objectives of this research is trying to improve impact evaluation within a particular government, this approach would involve working with them to identify an example of a good impact evaluation, then to find out how they managed to achieve that and then, explore whether their practices might be transferable. This approach suggests a fundamental shift in how the research would be done from researcher-led to intended user-led.

6 Examples of important research questions about impact evaluation and how they might be answered

To illustrate what these different ideas might look like in practice, Tables 3–5 set out some research questions, grouped in terms of the different aspects of an impact evaluation. While they are all genuine research questions, which could contribute to improving impact evaluation in and for development, they have also been chosen to illustrate different types of research approaches that could be used.

7 Conclusion

The development of a formal research agenda will require a consultative process of identifying those who might contribute or benefit in various ways to identify needs, priorities and opportunities. It needs sufficient resources. And, it needs the right combination of creative abrasion and interdisciplinary cooperation.

The range of possible research questions is large. The scope for fieldwork and subsequent uptake is also large. Researchers from a number of different disciplines will be needed to do this well. This interdisciplinary ‘creative abrasion’ can help to surface assumptions about evaluation which will add to the value of the research.

Increasing efforts at international collaboration, including special events around the International Year of Evaluation in 2015, could provide both resources, networking opportunities and impetus to formalise the research agenda and proceed to fund its implementation over a number of years.

Notes

[1] The process was conducted in Bolivia, Botswana, the DRC, India, Kenya, Lesotho, Mozambique, Namibia, Nicaragua, Papua New Guinea, Rwanda, South Africa, Tanzania, Thailand, Uganda and Uruguay.

[2] www.betterevaluation.org (accessed 30 July 2014).

[3] Most programme theories show how an intervention is expected to contribute to positive impacts. Negative programme theory shows how it might produce negative impacts. See http://betterevaluation.org/evaluation-options/negative_program_theory.

References

Beynon, P.; Chapoy, C.; Gaarder, M. and Masset, E. (2012) *What Difference does a Policy Brief Make?*, Full Report of an IDS, 3ie, Norad study, Brighton: IDS

Braskamp, L.A.; Brown, R.D. and Newman, D.L. (1982) ‘Studying Evaluation Utilization Through Simulations’, *Evaluation Review* 6.1: 114–26

Coryn, C.L.; Hattie, J.A.; Scriven, M. and Hartmann, D.J. (2007) ‘Models and Mechanisms for Evaluating Government-Funded Research. An International Comparison’, *American Journal of Evaluation* 28.4: 437–57

Coryn, C.L.; Noakes, L.A.; Westine, C.D. and Schröter, D.C. (2011) ‘A Systematic Review of Theory-Driven Evaluation Practice from 1990 to 2009’, *American Journal of Evaluation* 32.2: 199–226

Cousins, J.B. and Leithwood, K.A. (1986) ‘Current Empirical Research on Evaluation Utilization’, *Review of Educational Research* 56.3: 331–64

Cranston, P.; Beynon, P. and Rowley, J. (2013) *Two Sides of the Evaluation Coin: An Honest Narrative Co-Constructed by the Commissioner and the Contractor Concerning One Evaluation Experience*, Melbourne: BetterEvaluation, http://betterevaluation.org/resource/example/two-sides-coin (accessed 30 July 2014)

Douglas, J. (1976) *Investigative Social Research. Individual and Team Field Research*, Beverly Hills CA: Sage Publications

Elmore, R.F. (1978) 'Organizational Models of Social Program Implementation', *Public Policy* 26.2: 185

GAO (1990) *Case Study Evaluation*, Washington DC: Government Accounting Office

Jayawickrama, J. (2013) '"If They Can't Do Any Good, They Shouldn't Come": Northern Evaluators in Southern Realities', *Journal of Peacebuilding and Development* 8.2: 26–41

Luo, L.P. and Liu, L. (2014) 'Reflections on Conducting Evaluations for Rural Development Interventions in China', *Evaluation and Program Planning* 47: 1–8

Mark, M.M. and Henry, G.T. (2004) 'The Mechanisms and Outcomes of Evaluation Influence', *Evaluation* 10.1: 35–57

MOHSS/DSP (2010) *Proceedings. Stakeholders Workshop on 'Setting a National Evaluation and Research Agenda for HIV/AIDS in Namibia'*, Windhoek: Ministry of Health and Social Services, Directorate of Special Programs (MOHSS/DSP)

NAMc (2010) *Proceedings. Consultative Workshop on 'Developing a National Evaluation Agenda for HIV/AIDS in Thailand'*, Cha-Am: National AIDS Management Center

Oakden, J. (2013) *Evaluation Rubrics: How to Ensure Transparent and Clear Assessment that Respects Diverse Lines of Evidence*, Melbourne: BetterEvaluation, http://betterevaluation.org/sites/default/files/Evaluation%20rubrics.pdf (accessed 5 August 2014)

OEDC-DAC (2010) *Glossary of Key Terms in Evaluation and Results Based Management*, Paris: OEDC, www.oecd.org/development/peer-reviews/2754804.pdf (accessed 5 August 2014)

Patton, M.Q.; Grimes, P.S.; Guthrie K.M.; Brennan, N.J.; French, B.D. and Blyth, D.A. (1975) *In Search of Impact: An Analysis of the Utilization of Federal Health Evaluation Research*, Minneapolis and St Paul: University of Minnesota

Pawson, R. (2002) *Evidence-Based Policy: The Promise of Realist Synthesis*, London: Sage Publications

Pawson, R. and Tilley, N. (1997) *Realist Evaluation*, London: Sage Publications

Peersman, G. (2010) *A National Evaluation Agenda for HIV. UNAIDS Monitoring and Evaluation Fundamentals*, Geneva: UNAIDS

Perrin, B. (2002) *Implementing the Vision: Addressing Challenges to Results-Focused Management and Budgeting*, in OECD meeting Implementation Challenges in Results Focused Management and Budgeting, 11–12 February 2002, Paris

Reed, E. and Moriaru, J. (2010) *State of Evaluation: Evaluation Practice and Capacity in the Nonprofit Sector*, Washington DC: Innovation Network

Richard P.; Sternin, J. and Sternin, M. (2010) *The Power of Positive Deviance: How Unlikely Innovators Solve the World's Toughest Problems*, Cambridge MA: Harvard Business Press

Rogers, P.J. and Hough, G.I. (1995) 'Improving the Effectiveness of Evaluations: Making the Link to Organizational Theory', *Evaluation and Program Planning* 18.4: 321–32

Rogers, P. and McDonald, B. (2001) *Impact Evaluation Research Project*, Melbourne: Department of Natural Resources and Environment

Rogers, P.J.; Petrosino, A.; Huebner, T.A. and Hacsi, T.A. (2000) 'Program Theory Evaluation: Practice, Promise, and Problems', *New Directions for Evaluation* 2000.87: 5–13

Schwarz, C. and Struhkamp, G. (2007) 'Does Evaluation Build or Destroy Trust? Insights from Case Studies on Evaluation in Higher Education Reform', *Evaluation* 13.3: 323–39

Scriven, M. and Rogers, P.J. (1995) 'Stories of Evaluation', panel presentation at the 1995 International Evaluation Conference on Evaluation for a New Century: A Global Perspective, American Evaluation Association co-sponsored with the Canadian Evaluation Society, Vancouver, British Columbia, Canada, November 1995

Shulha, L.M. and Cousins, J.B. (1997) 'Evaluation Use: Theory, Research, and Practice Since 1986', *American Journal of Evaluation* 18.3: 195–208

Smith, J. and Sweetman, A. (2009) 'Putting the Evidence into Evidence-Based Policy', in Productivity Commission (eds), *Strengthening Evidence-based Policy in the Australian Federation, Canberra, 17–18 August 2009, Volume 1: Roundtable Proceedings*, Canberra: Australian Government

USAID (2011) *USAID Evaluation Policy*, Washington DC: USAID, www.usaid.gov/sites/default/files/documents/2151/USAIDEvaluationPolicy.pdf (accessed 5 August 2014)

Valovirta, V. (2002) 'Evaluation Utilization as Argumentation', *Evaluation* 8.1: 60–80

Van Maanen, J. (2011) *Tales of the Field: On Writing*

Ethnography, 2nd ed., Chicago and London: University of Chicago Press

White, H. (2009) *Theory-based Impact Evaluation: Principles and Practice*, 3ie Working Paper 3, www.3ieimpact.org/media/filer_public/2012/05/07/Working_Paper_3.pdf (accessed 5 August 2014)

Wildschut, L.P. (2014) 'Theory-Based Evaluation, Logic Modelling and the Experience of South African Non-Governmental Organisations', PhD dissertation, South Africa: Stellenbosch University